UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

Departamento de Química Física I

TESIS DOCTORAL

Evaluación de potenciales de plegamiento de proteínas con algoritmos genéticos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

David de Sancho Sánchez

Director

Antonio Rey Gayo

Madrid, 2018

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

DEPARTAMENTO DE QUÍMICA FÍSICA I

EVALUACIÓN DE POTENCIALES DE PLEGAMIENTO DE PROTEÍNAS CON ALGORITMOS GENÉTICOS

Memoria que presenta para optar al grado de Doctor en Ciencias Químicas

David de Sancho Sánchez

Director: Dr. Antonio Rey Gayo

Madrid, 2007

Agradecimientos y dedicatoria

Esta Tesis Doctoral ha sido posible gracias a la financiación por parte del Ministerio de Ciencia y Tecnología, ahora de Educación y Ciencia, del proyecto BQU2002-04626-C02-01, y de la beca BES-2003-3099 de la que he sido concesionario.

La primera persona a la que quiero dar las gracias es Antonio Rey, el director de esta Tesis Doctoral. Le agradezco su seriedad y su rigor, su capacidad para comprender y sus ganas de enseñarlo todo. En fin, Antonio, te doy las gracias porque eres el mejor de los maestros.

En segundo lugar, quiero dar las gracias a Lidia Prieto y a María Larriva. Lidia era la chica rubia que hacía el proyecto con Antonio cuando entré en el grupo. Ahora es una amiga sin par. María vino un poco más tarde, pero siempre ha sido tan generosa como para seguirme la corriente elucubrando sobre efectos hidrófobos. Las dos me han cuidado con una delicadeza extraordinaria. Mi agradecimiento por ese cuidado no cabe aquí, y quizás tampoco en las lágrimas con que me despediré de ellas. Chicas, de verdad, sois inigualables.

Esta Tesis se ha realizado en el Departamento de Química Física I, a cuyos miembros quiero dar las gracias. Muy especialmente recuerdo a los becarios Eduardo Sanz, Andrés Guerrero, Francisco Javier Martínez —Willy—, Eduardo Pérez, Elena del Corro, Yolanda Sánchez y César Mediavilla, con quienes mejor me lo he pasado y de quienes más he aprendido. Por cierto, que mi ala del departamento está limpia como una patena gracias a Paqui, a la que mando un fuerte abrazo. También quiero dar las gracias a otros compañeros investigadores: Alfender, María José Dávila, Laura Miranda, María... También a los bioquímicos, especialmente a Florian Baumgart, Elías Herrero y Jorge Alegre.

En 2004, Lidia y yo viajamos a Gerona para realizar nuestros cursos de doctorado. Allí vivimos una experiencia maravillosa, de la que lo mejor fue trabar conocimiento con un grupo de personas realmente increíble. Gracias todos los profesores y compañeros de aquel curso, muy especialmente a Julia Contreras, Merche Alonso, Nuria González,

Ainara Nova, Judit Durá, Raquel Ríos, y muchos más. En ese curso de doctorado también conocí a Natal Kanaan. Natal me cogió de la mano y abrió las ventanas de mi casa. Osita, muchas gracias.

En los últimos años he tenido la oportunidad de viajar a Cardiff y Amsterdam, para realizar sendos cursos financiados por la Unión Europea. Quiero dar las gracias por esa oportunidad. Gracias también a compañeros y profesores, a Luana Sucupira, Daniella Botelho y Martin Burke, de Cardiff, y a Alexis Rutherford, Iain Johnston, Richard Matthews y Gleb Solomentsev, de Amsterdam. También en Zaragoza, adonde he ido para realizar cursos y asistir a congresos, nos han tratado siempre muy bien. Gracias a los miembros del BIFI —y especialmente a Javier Sancho— por sus muchas atenciones.

Todas estas personas han sido importantes en el desarrollo de la Tesis, digamos, por dentro. Pero también hay muchas personas que me han acompañado todos estos años por fuera. Nacho Cerezo, Guillermo Gea, David Rodríguez, Andrés Arregui, Gonzalo Givaja, Montaña Lindo, Teresa Ballester; en fin, mis amigos: gracias siempre por quererme. Laura, Jaime, Paula, Anna, Carolina, Beatriz, Paloma, Ana, Sarinha, Angélica, Eugenia, Sara, Elisa, Judit, Fran... Gracias a todos.

También quiero dar las gracias a mi hermano Ignacio, y a la familia de mi tío Manolo.

Dos personas que me han ayudado a perderle el miedo a mis propios abismos son Enrique y Cecilia, mis maestros de yoga, y Sergio Larriera, mi psicoanalista. Con su trabajo hacen de mi vida algo mejor —algo posible—. Por eso, muchas gracias.

Dentro de esa clasificación tan poco válida —dentro de la Tesis, fuera— hay una persona que excede los márgenes especialmente. Se trata de María Muñoz Caffarel. En fin, María, por ser tan especial, por cometer la dulcísima insensatez de compartir tus días con los míos, por hacerme tan feliz, gracias.

Finalmente, quiero dedicar esta Tesis Doctoral a mi madre, Ángeles Sánchez Molano, la persona que más se ha esforzado en que haya podido llegar hasta aquí.

Índice general

Model makers are storytellers. Stories can be involved and detailed, or, like simple models, they can describe just the essentials. The art of model building is in recognizing what is essential.

Los modelizadores son contadores de cuentos. Los cuentos pueden ser complejos y detallados, o, como los modelos sencillos, pueden describir sólo lo esencial. El arte de modelizar reside en descubrir qué es lo esencial.

Ken A. Dill & Sarina Bromberg, "Molecular Driving Forces"

Capítulo 1

Introducción

En esta memoria se resume la investigación realizada con el objeto de evaluar potenciales de interacción para el plegamiento de proteínas. Hemos llevado a cabo esta evaluación utilizando un método computacional de minimización basado en algoritmos genéticos. En este Capítulo introducimos los conceptos más importantes en el campo del plegamiento de proteínas, de los que haremos uso a lo largo de esta memoria. El proceso de plegamiento se puede abordar desde muy distintas aproximaciones, tanto experimentales como teóricas. Hacemos especial hincapié en los métodos de simulación molecular más utilizados para el estudio teórico del plegamiento. Entre ellos, los algoritmos genéticos tienen características particulares que los hacen apropiados para nuestro estudio. También de ellas damos cuenta en este Capítulo. Finalmente, describimos la organización de esta memoria.

1.1. Estructura de proteínas y plegamiento

Las proteínas son heteropolímeros lineales formados por veinte tipos de aminoácidos naturales[1]. En un aminoácido, un grupo amino y un grupo carboxilo se encuentran uni-

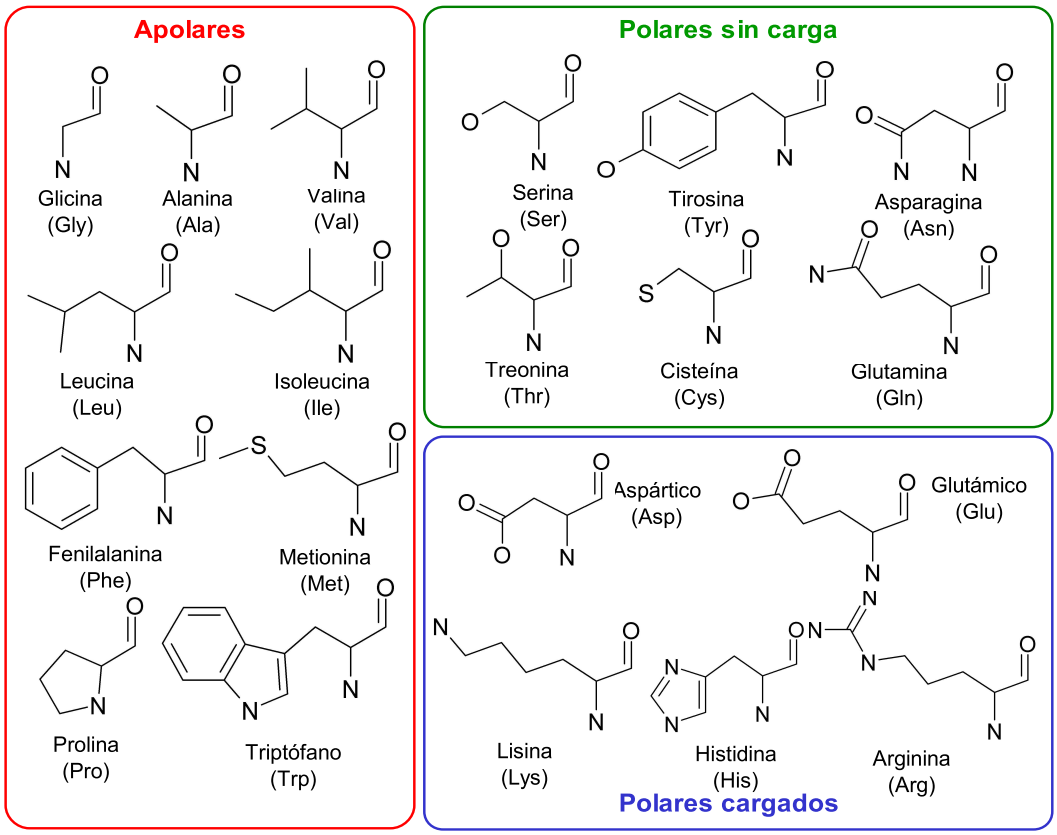


Figura 1.1: Clasificación de los aminoácidos naturales que aparecen en las proteínas en función de la naturaleza de su cadena lateral.

dos a un átomo de carbono, generalmente asimétrico, llamado carbono-α. Los distintos aminoácidos se distinguen por su cadena lateral, también unida al carbono-α. En la Figura 1.1 representamos los aminoácidos naturales que forman las proteínas, clasificados en función de la naturaleza de su cadena lateral como apolares, polares sin carga y polares cargados.

En una proteína, los aminoácidos se unen entre sí a través de enlaces peptídicos[1]. El enlace peptídico es de tipo amida, y se establece entre el grupo carboxilo de un aminoácido y el grupo amino del siguiente en la secuencia. Es un enlace parcialmente doble, por lo que tiene una geometría prácticamente plana. Debido a esta geometría, en una cadena de residuos de aminoácido se define una sucesión de planos peptídicos. En la Figura 1.2 mostramos los planos correspondientes a los dos enlaces peptídicos que

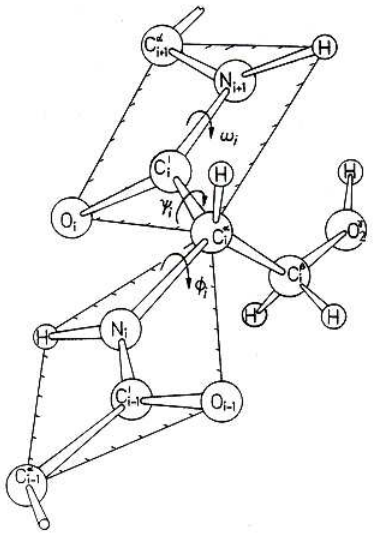


Figura 1.2: Fragmento de una cadena polipeptídica: mostramos los ángulos de Ramachandran ϕ y ψ para un residuo i de serina, y el ángulo ω del plano peptídico entre los residuos i e $(i+1)$.

forma un residuo de serina en un fragmento de cadena polipeptídica. Aunque las desviaciones de la planalidad suelen ser pequeñas, para la rotación de cada enlace peptídico de una cadena se define el ángulo de torsión ω. En proteínas, la forma *trans* ($\omega = 180°$) predomina en relación 1000:1 sobre la forma *cis* ($\omega = 0°$) por repulsiones estéricas entre átomos de aminoácidos vecinos. Además, los planos peptídicos pueden rotar sobre los ángulos diedros ϕ y ψ, llamados ángulos de Ramachandran, que también mostramos en la Figura 1.2. Los valores que pueden adoptar los ángulos de Ramachandran también están restringidos debido a las colisiones estéricas. Estas restricciones en ϕ y ψ permiten definir para cada clase de aminoácido el denominado mapa de Ramachandran[2], cuyas distintas zonas corresponden a los tipos estructurales accesibles para ese aminoácido.

Para el esqueleto de una proteína con N aminoácidos en su secuencia, los grados de libertad conformacionales son los $(N-1)$ conjuntos de ángulos $\{\omega, \phi, \psi\}$. A pesar de las restricciones en los valores de ϕ, ψ y, sobre todo, ω, la cadena polipeptídica tiene una gran libertad conformacional. También las cadenas laterales de la mayoría de los

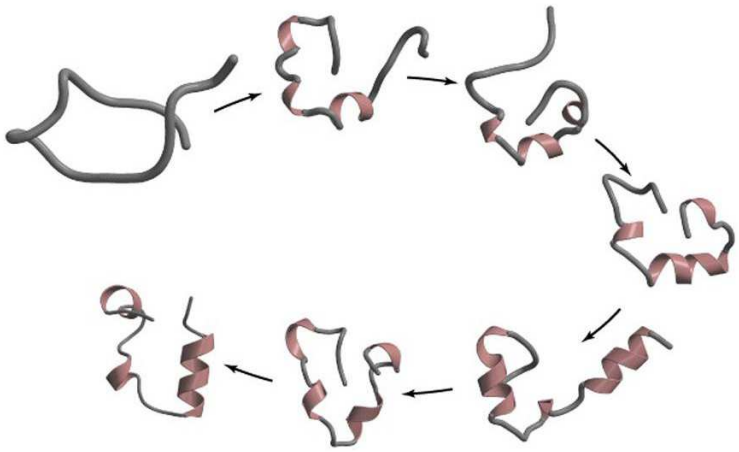


Figura 1.3: Representación simplificada del proceso de plegamiento de una proteína.

residuos que forman una proteína pueden adquirir diferentes conformaciones. Por tanto, considerando el esqueleto peptídico y las cadenas laterales, una proteína tiene una gran flexibilidad.

Debido al gran número de grados de libertad, una vez sintetizada una proteína puede adquirir muchas conformaciones diferentes. Sin embargo, se ha observado que, en medio acuoso, las proteínas adquieren reproduciblemente y en muy poco tiempo una estructura tridimensional concreta[3]. Esta estructura, necesaria para que la proteína realice la función que tiene asignada en la naturaleza, se denomina conformación nativa. El plegamiento, representado esquemáticamente en la Figura 1.3, es el proceso que permite a una cadena polipeptídica adquirir esta conformación.

1.2. La hipótesis termodinámica de Anfinsen

A partir de una serie de estudios experimentales, se ha conocido la capacidad de las proteínas de alcanzar el estado nativo a partir de su forma desplegada o estado desnaturalizado. Estos estudios alcanzaron su madurez con los trabajos de Christian B. Anfinsen sobre la renaturalización de la ribonucleasa[4,5]. En sus experimentos, Anfinsen

sometía a la ribonucleasa a un tratamiento desnaturalizante con urea 8 M. Así obtenía una mezcla de productos que conservaba un 1 % de la actividad de la enzima en su forma nativa. Al retirar la urea, la enzima recuperaba el 100 % de su actividad. A partir de estas observaciones, Anfinsen concluyó que la renaturalización de la enzima era un proceso enteramente dirigido por la energía libre. Al recuperar su conformación nativa, la proteína alcanzaba el mínimo de energía libre del sistema.

Los resultados obtenidos por Anfinsen le llevaron a enunciar la llamada *hipótesis termodinámica*, que establece que

> *la estructura tridimensional de la conformación nativa de una proteína en su medio fisiológico habitual (disolvente, pH, fuerza iónica, presencia de otros componentes como iones metálicos o grupos prostéticos, temperatura, etc.) es aquella para la cual la energía libre de Gibbs de todo el sistema es mínima*[4].

De acuerdo con esta hipótesis, la conformación nativa está dictada por la totalidad de sus interacciones. En gran medida, estas interacciones vienen determinadas por la secuencia de aminoácidos de la proteína.

1.3. Paradigmas para el plegamiento

Un problema que emerge del elevado número de grados de libertad de una proteína es cómo se busca la conformación de mínima energía de la que hablaba Anfinsen. Para entender mejor el alcance de este problema, podemos valernos del concepto de superficie o paisaje de energía ("energy landscape").

En este caso concreto, definimos una superficie de energía como una representación de la energía libre del sistema en función de los grados de libertad de una proteína[6]. Estos grados de libertad pueden ser, por ejemplo, los valores de los ángulos diedros ϕ y ψ para todos los residuos de la cadena. Cada conformación de la proteína corresponde

a un punto en la superficie de energía. Para especificar la posición de este punto, sería necesaria una representación de muchas dimensiones, tantas como grados de libertad tiene la cadena. En la Figura 1.4 mostramos una serie de superficies de energía para el plegamiento representadas de manera muy idealizada. En ellas, únicamente aparecen dos coordenadas hipotéticas. Cada conformación se define por sus valores para estas dos coordenadas. En el eje vertical de estas superficies de energía se representa la "energía libre interna" de cada conformación, que tiene su mínimo para la conformación nativa. La "energía libre interna" contiene todas las contribuciones a la energía excepto la entropía conformacional, que es función de la degeneración de cada nivel energético[7]. Para cada valor de la energía libre interna, la entropía conformacional está relacionada con la anchura de la superficie.

Podemos utilizar superficies de energía como las de la Figura 1.4 para comprender los distintos paradigmas que se han utilizado para comprender el proceso de plegamiento. La superficie que mostramos en la Figura 1.4 (a) corresponde al modelo de búsqueda aleatoria. En este caso, el paisaje de energía es una superficie plana, excepto por el mínimo que corresponde a la conformación nativa (N). Esta superficie supone que una cadena polipeptídica tendría que buscar al azar su estructura nativa para poder plegarse correctamente. Si consideramos el número de grados de libertad de una proteína, para recorrer todas las conformaciones accesibles se requeriría un periodo de tiempo mayor que la edad del Universo. Sin embargo, como hemos comentado, las proteínas tardan en plegarse muy poco tiempo, entre milisegundos y segundos[8]. Por tanto, es imposible que el proceso de plegamiento de una proteína sea aleatorio.

Fue Cyrus Levinthal quien formuló y dio una primera solución a este problema, que recibió el nombre de "paradoja de Levinthal". Esta solución consistiría en que existiesen caminos para el plegamiento[9,10], como sucede en el caso de las reacciones unimoleculares sencillas[11]. Las reacciones de este tipo están gobernadas por el paso del sistema por un

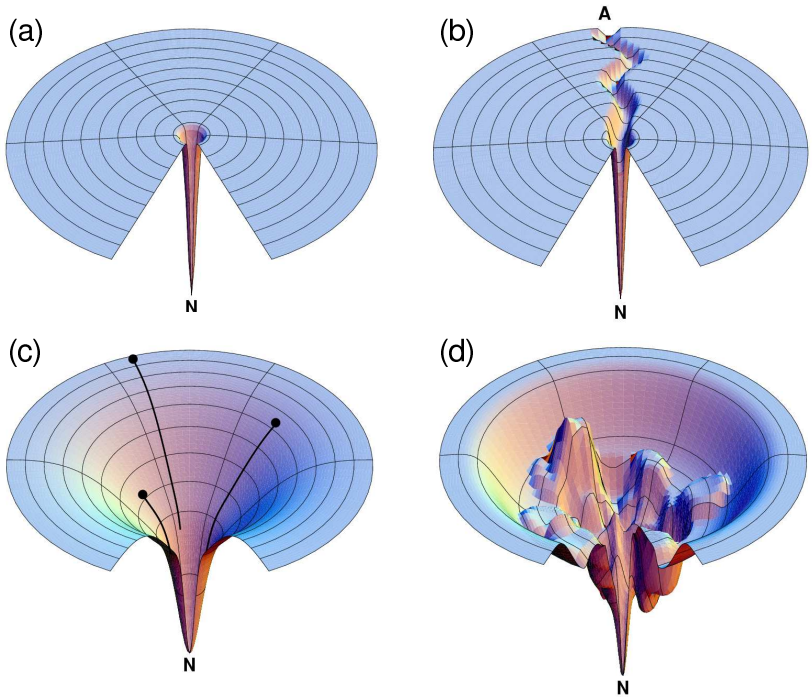


Figura 1.4: Los distintos paradigmas para el proceso de plegamiento a través de superficies de energía[6]. (a) Búsqueda al azar o "paradoja de Levinthal", (b) camino de plegamiento, (c) embudo liso, y (d) embudo rugoso. En todos ellos, el eje vertical corresponde a la "energía libre interna", la anchura, a la entropía conformacional y N es el estado nativo.

estado de transición. De acuerdo con esta analogía, podríamos representar la superficie de energía como en la Figura 1.4 (b). El proceso de plegamiento podría verse como una reacción en la que un reactivo, la proteína desplegada (A), se convierte en un producto, el estado nativo (N)[11]. La reacción transcurriría por un camino bien trazado en la superficie de energía, constituido por eventos sucesivos que dirigirían el plegamiento[6]. El problema de este modelo es que no está clara la base física de esta serie de eventos en el plegamiento. Además, no tiene en cuenta la heterogeneidad del estado desnaturalizado.

A partir de una serie de trabajos sobre modelos simplificados[12–19], se llegó a un nuevo paradigma del proceso de plegamiento, el denominado embudo de plegamiento. De acuerdo con este paradigma, no habría caminos únicos con un solo estado de transición,

sino que la proteína podría plegarse de diferentes maneras, que convergerían en su camino hacia la conformación nativa[20]. Los trabajos sobre cinética del plegamiento del grupo de Alan Fersht en seguida dieron respaldo experimental a este nuevo modelo[21]. En la Figura 1.4 (c) mostramos un embudo ideal de superficie lisa que permite comprender este nuevo paradigma para el plegamiento. En el embudo, diferentes conformaciones desplegadas muy alejadas entre sí convergen en su camino hacia la conformación nativa, pasando por un grupo menos numeroso de estructuras compactas. La forma de embudo de la superficie de energía es, de acuerdo con este modelo, responsable de la robustez del plegamiento[22].

Este embudo de superficie lisa es muy improbable si pensamos en la complejidad de un sistema como una proteína. Un modelo más realista del embudo de plegamiento es el que se muestra en la Figura 1.4 (d). Como en el caso anterior, al aproximarse la proteína a su conformación nativa disminuye progresivamente la anchura de la superficie. Sin embargo, en este caso se aprecian rugosidades en la superficie del embudo. Habría por tanto un mínimo absoluto, pero también muchos mínimos locales. Estos mínimos locales suponen trampas en las que la proteína puede residir temporalmente durante el proceso de plegamiento[23].

1.4. Principales contribuciones energéticas al plegamiento de proteínas

Hemos comentado una serie de paradigmas para el proceso de plegamiento. Entre ellos, el más utilizado actualmente es el último que hemos presentado, el del embudo rugoso de la Figura 1.4 (d). Su rugosidad se debe a las interacciones estabilizantes que se producen entre residuos de la proteína, y entre los residuos y el disolvente[20]. Las interacciones que estabilizan la conformación nativa de la proteína se oponen al efecto de

la entropía conformacional. El efecto entrópico se debe, sencillamente, a que para un heteropolímero lineal como una proteína hay muchas más conformaciones posibles en estados desplegados que en estructuras compactas como la nativa[24].

Las interacciones que contribuyen de manera más decisiva al plegamiento de proteínas son los enlaces de hidrógeno y el efecto hidrófobo. También las interacciones electrostáticas estabilizan, aunque de manera menos importante, la conformación nativa. Para estudiar la estabilidad de las proteínas, suelen utilizarse experimentos de calorimetría diferencial de barrido[25]. A partir de los resultados de esta técnica, se puede calcular la energía libre del desplegamiento, es decir, del paso del estado nativo al estado desnaturalizado. El balance de contribuciones resultante en el caso del desplegamiento es:

$$\Delta G \approx \Delta G_{\text{efecto hidrófobo}} + \Delta G_{\text{enlace de hidrógeno}} + \Delta G_{\text{electrostática}} - T\Delta S_{\text{conformacional}} \quad (1.1)$$

En el intervalo de temperaturas de interés, se ha determinado que ΔG para este proceso es aproximadamente 20-60 kJ/mol[25]. Por tanto, la estructura nativa de una proteína es sólo marginalmente más estable que su forma desnaturalizada. Dado lo ajustado que es el balance de contribuciones, puede decirse que todas las interacciones —de las que hablamos a continuación— son significativas en el plegamiento.

1. Enlaces de hidrógeno: En una proteína todos los residuos se encuentran, en promedio, formando uno o más enlaces de hidrógeno[26]. Los más abundantes son los que se establecen entre grupos del esqueleto peptídico. En este caso, el enlace de hidrógeno se forma entre el átomo de nitrógeno unido a hidrógeno de un grupo amino (donador), y el átomo de oxígeno de un grupo carbonilo (aceptor). Los enlaces de hidrógeno del esqueleto estabilizan los principales tipos de estructura secundaria, las hélices α y las láminas β. Además, se pueden formar enlaces de hidrógeno entre

un grupo del esqueleto y otro de una cadena lateral, y entre cadenas laterales. A pesar de su ubicuidad en proteínas, hay desacuerdo en cuanto a la magnitud de la contribución de los enlaces de hidrógeno a la energía global[24,27–29].

2. Interacciones hidrófobas: A las interacciones hidrófobas se les atribuye generalmente el papel de fuerza directora del plegamiento[24,30]. En gran medida, la contribución más importante de estas interacciones es entrópica. Cuando una proteína en disolución se encuentra en su estado desplegado, un gran número de moléculas de agua están "secuestradas" por su interacción con la cadena. En cambio, cuando la proteína adquiere una conformación compacta, el número de moléculas de agua que interacciona con la proteína es mucho menor. Por tanto, la entropía del agua en el estado plegado de la proteína es mucho mayor que en su estado desnaturalizado. Esta diferencia de entropía entre los dos estados explica la tendencia de una proteína a colapsar[24]. Al producirse el colapso, los residuos de la proteína pasan de un entorno polar a otro apolar. Así, forman una región densamente empaquetada, protegida del disolvente. No todos los aminoácidos tienen la misma tendencia a pasar de un entorno polar a otro apolar. La tendencia de cada aminoácido a aparecer en el interior o en el exterior de las proteínas puede relacionarse con su posición en una escala de hidrofobia[31,32]. En la siguiente sección hablaremos de cómo estas diferencias en la hidrofobia pueden capturarse, utilizando métodos basados en la termodinámica estadística, para diseñar potenciales de interacción para proteínas.

3. Interacciones electrostáticas: Las interacciones electrostáticas se forman en las proteínas debido a la presencia de cargas con distinto signo en las cadenas laterales de los residuos de aminoácido. Se ha observado que las interacciones de este tipo en el interior de una proteína, los llamados puentes salinos, son generalmente desestabilizantes[33]. Por el contrario, las que se forman en la superficie tienen un efecto

estabilizante, que puede verse afectado por la constante dieléctrica del medio. Así, la mayoría de grupos cargados que forman interacciones electostáticas se localiza en la superficie de las proteínas[25]. Su contribución al plegamiento es, en todo caso, minoritaria. Sin embargo, un síntoma de su importancia es que las interacciones electrostáticas específicas pueden ser utilizadas por la Naturaleza para optimizar la estabilidad en el "diseño" de determinado tipo de proteínas[34].

1.5. Modelización del plegamiento

El plegamiento de proteínas puede ser abordado mediante aproximaciones tanto experimentales como teóricas. Las aproximaciones teóricas pueden dividirse, a su vez, en dos grupos, dependiendo del problema que se intente resolver. En un primer grupo de aproximaciones teóricas, se trata de predecir la estructura tridimensional de una proteína a partir de su secuencia. En el segundo grupo, es el propio proceso de plegamiento lo que se intenta comprender. Algunos métodos que se utilizan para la predicción de estructura son, por ejemplo, el "threading" o enhebrado —del que hablaremos más adelante— o el modelado por homología[35,36]. Por otra parte, para estudiar el proceso de plegamiento en sí, existen distintas aproximaciones posibles, que trataremos a continuación. En todas ellas hay dos elementos necesarios: una descripción de la estructura de la proteína y sus interacciones, y un algoritmo de muestreo conformacional eficaz[37]. Estos elementos suelen estar íntimamente ligados entre sí.

Entre las aproximaciones teóricas, los estudios más detallados son aquellos en que se intenta realizar una simulación atomística del proceso de plegamiento desde una estructura arbitraria[38]. Además de considerar todos los átomos de la proteína, suelen tenerse en cuenta explícitamente las interacciones con el disolvente. Por ello, se consideran varios millares de moléculas de agua en el sistema simulado. Proteína y disolvente se

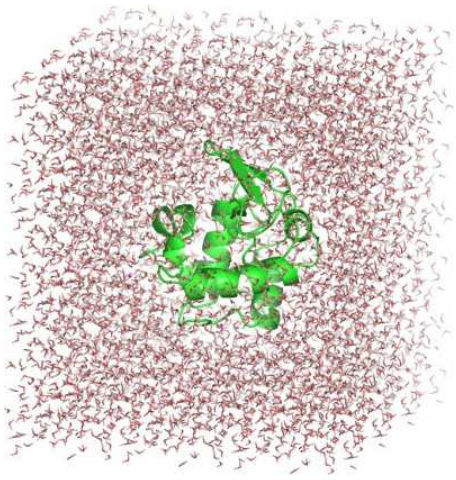


Figura 1.5: Sistema formado por una molécula de lisozima (en verde) dentro de una caja de agua para realizar una simulación atomística por dinámica molecular de sus estados desplegados.

introducen en una caja convenientemente diseñada, como la que mostramos en la Figura 1.5 para una simulación atomística de la lisozima. La simulación de un gran número de moléculas de agua para estudiar propiedades de una proteína supone un problema de eficiencia en este tipo de métodos. Una alternativa menos costosa computacionalmente es la utilización de modelos implícitos para el agua. En los estudios atomísticos del plegamiento, la energía global suele calcularse como suma de contribuciones. Los términos de esa suma son dependientes de longitudes y ángulos de enlace, ángulos de torsión, interacciones de van der Waals e interacciones electrostáticas[38]. Estas contribuciones, convenientemente parametrizadas a partir de datos experimentales y resultados de cálculos *ab initio*, se engloban en campos de fuerza de mecánica molecular, como AMBER[39], CHARMM[40], GROMOS[41] y OPLS[42].

El método de simulación que se utiliza mayoritariamente con los modelos atomísticos es la dinámica molecular. El fundamento del método de dinámica molecular es la resolución de las ecuaciones de Newton para obtener propiedades de equilibrio y de transporte del sistema[43–45]. Con una descripción de las interacciones como la de los campos de fuerza, las ecuaciones de movimiento pueden ser resueltas para cada átomo

de la proteína. Así, en teoría, se puede seguir la trayectoria del plegamiento e identificar el estado nativo en el mínimo energético. El problema de este tipo de método para la simulación del plegamiento lo encontramos en que, con la capacidad de cálculo actual, para sistemas con un gran número de átomos como el que hemos descrito, únicamente pueden alcanzarse tiempos de simulación de varios nanosegundos[46,47]. Como hemos comentado, la escala de tiempo en que una proteína se pliega es de milisegundos a segundos. De modo que el plegamiento excede las capacidades de la dinámica molecular, salvo quizás para los denominados plegadores ultrarrápidos ("ultra-fast folders")[48]. Por ello, muchas veces se simula a elevada temperatura el desplegamiento, cuyo cálculo resulta mucho más rápido que el plegamiento, aunque no proporcione con seguridad la misma información[38].

La alternativa a la modelización atomística de una proteína es la modelización "coarse-grained" o de grano grueso[49–51]. En los modelos de grano grueso más simplificados, como los que mostramos en la Figura 1.6, cada aminoácido de la proteína está representado por un solo centro de interacción. Además, en el modelo, las posiciones de estos centros están discretizadas en los nudos de una red. Normalmente las interacciones se representan de una manera muy simplificada. Por ejemplo, en el modelo HP[16] sólo se consideran dos tipos de aminoácidos, hidrófobos (H) y polares (P). A pesar de su extrema sencillez, el estudio de este tipo de modelos ha permitido obtener conclusiones de gran importancia para el estudio del plegamiento[16].

Otros modelos de grano grueso son aquellos que consideran un nivel de detalle intermedio[50,51]. En este caso se puede representar un mayor número de centros de interacción por residuo. En el modelo, estos centros pueden ocupar posiciones de una red de alta resolución. También pueden utilizarse modelos fuera de red. En estos modelos, a pesar de utilizarse un número pequeño de centros de interacción por residuo de una proteína, el muestreo del espacio conformacional es mucho más costoso. Mostramos dos

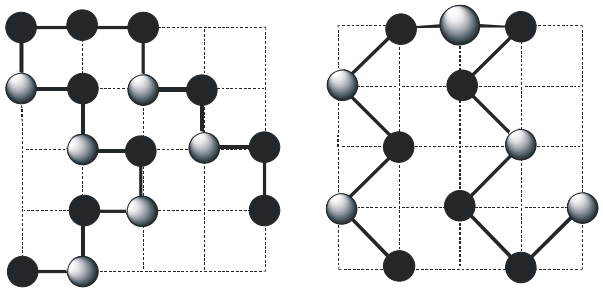


Figura 1.6: Modelos en red de baja resolución para la simulación del plegamiento. A la izquierda, red bidimensional cuadrada. A la derecha, red tridimensional cúbica centrada en las caras. En ambas se utiliza el modelo HP. Los colores blanco y negro manifiestan los dos tipos de residuo, hidrófobos y polares.

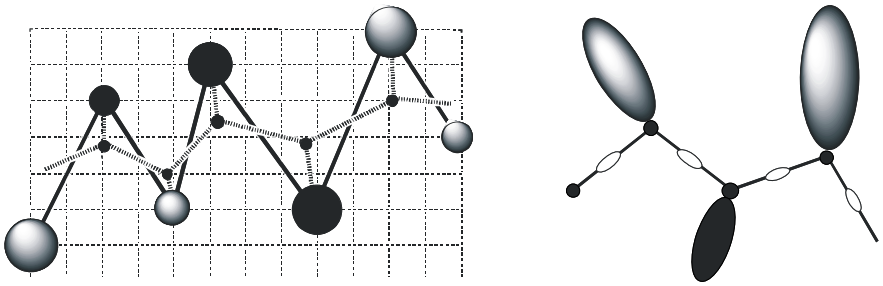


Figura 1.7: Modelos de resolución media para simulación del plegamiento. A la izquierda, modelo SICHO, con dos centros por residuo: centros de las cadenas laterales, en las posiciones de una red cúbica, y carbonos-α, fuera de red. A la derecha, modelo UNRES, con tres centros por residuo, todos fuera de red: carbonos-α (círculos negros), centros de los enlaces peptídicos (en blanco) y centros de las cadenas laterales (elipsoides).

ejemplos de este tipo de representación de la proteína, los modelos SICHO[52] y UNRES[53], en la Figura 1.7.

Con los modelos de resolución intermedia, para el cálculo de la energía suelen utilizarse potenciales empíricos[54,55]. Cuando se usa este tipo de potencial, la energía global se define como la suma ponderada de términos correspondientes a interacciones de corto y largo alcance, genéricas y dependientes de secuencia. Muchas veces, para el diseño de este tipo de potenciales se parte de la información estructural de proteínas obtenida experimentalmente, que se almacena en el Protein Data Bank (PDB)[56,57]. Dado que se utiliza información experimental para su obtención, estos potenciales suelen

denominarse "knowledge-based" o basados en estructuras.

Tanto para la simulación con los modelos en red más complejos como para los modelos fuera de red, se utiliza casi invariablemente el método de Monte Carlo. Este método permite obtener propiedades promedio del sistema de interés realizando un "muestreo de importancia" sobre su superficie de energía[43,44]. En un muestreo de importancia, en vez de recorrer homogéneamente todos los estados en los que puede encontrarse el sistema, hay una mayor probabilidad de contabilizar aquellos estados que son energéticamente más favorables. El problema de este método en el caso del plegamiento de proteínas es que, a bajas temperaturas, la proteína puede quedar atrapada en mínimos locales de la superficie de plegamiento[58–62]. Por ello, a menudo se utilizan métodos más sofisticados en los que se visitan estados accesibles a distintas temperaturas[63]. Así, durante la simulación se puede caracterizar tanto el estado desplegado como el estado nativo.

1.6. Algoritmos genéticos y plegamiento de proteínas

Además de los métodos de simulación más tradicionales, Monte Carlo y dinámica molecular, otras herramientas computacionales han contribuido al estudio del plegamiento de proteínas. Entre ellas, son de especial interés los algoritmos genéticos. Este tipo de técnica se basa en los mecanismos de la evolución natural para alcanzar soluciones óptimas —o próximas a la óptima— de un problema de búsqueda[64–66]. Se trata, a diferencia de la dinámica molecular y el método de Monte Carlo, de un método de optimización. Para buscar el óptimo de una determinada función de interés, la llamada *función "fitness"* o función de mérito, las soluciones posibles del problema de búsqueda son transformadas en cadenas de variables, denominadas *cromosomas*. Como en la Naturaleza, los cromosomas del algoritmo genético son trasformados a lo largo de sucesivas *generaciones*, lo

que permite que mejore la información que contienen. Los cambios en los cromosomas se llevan a cabo mediante una serie de *operadores genéticos*. Estos operadores, que se describirán en detalle en el siguiente Capítulo, son el de *replicación*, "*crossover*" o *entrecruzamiento* y *mutación*. Estos operadores permiten que se transfiera información entre ellos o que se incorpore nueva información en la población. Así, a lo largo de una serie de generaciones, este tipo de algoritmo alcanza su solución óptima.

En el campo de la química de proteínas, se han desarrollado implementaciones de algoritmos genéticos para búsqueda conformacional, "docking", diseño molecular, diseño de receptores, etc[67]. En el ámbito del plegamiento, las aproximaciones en las que se han utilizado este tipo de programas son muy diversas[68]: se han empleado tanto algoritmos genéticos en solitario como combinados con otros métodos de muestreo, como Monte Carlo[60,69,70], búsqueda Tabú[71] o técnicas de optimización local[72]. En cuanto a la representación de la proteína, se han llevado a cabo estudios que utilizan desde modelos reducidos, en red[70,73] o fuera de red[74], hasta modelos con detalle atómico[75,76]. Por último, se ha tratado de optimizar funciones *fitness* de muy distinto grado de complejidad: modelos de interacción tan simplificados como el citado HP[70,71,77]; potenciales aditivos que contabilizan únicamente algunas contribuciones a la estabilidad, como los enlaces de hidrógeno, los puentes disulfuro o la compactibilidad de la estructura[72,78,79]; o campos de fuerza empíricos incluidos en paquetes de cálculo como AMBER[75].

1.7. Objetivos y organización de esta Tesis

Como hemos comentado, los potenciales basados en estructuras parten de los datos sobre estructura de proteínas obtenidos experimentalmente y depositados en el PDB[56,57]. Estos datos son utilizados de una u otra manera para obtener potenciales para los distintos tipos de interacción —especialmente, enlaces de hidrógeno e interacciones hidrófobas—.

Esta aproximación ha sido utilizada por un gran número de grupos de investigación para generar sus propias funciones energéticas con las que estudiar el plegamiento[80–96].

Después de su elaboración, los potenciales de interacción para el plegamiento de proteínas deben ser evaluados. Un test que suelen realizar los grupos responsables de la obtención de estos potenciales es estudiar su capacidad para definir un mínimo energético en la conformación nativa para un conjunto de proteínas. Los métodos más utilizados con este fin están basados en "decoys" o estructuras quimera[97]. Estas quimeras se generan mediante el método denominado "threading" o enhebrado de la secuencia de una proteína sobre el molde de la estructura nativa de otras proteínas. Así se obtienen plegamientos alternativos al nativo para una determinada secuencia. A continuación, se calcula la energía de la conformación nativa y la de todas las quimeras. Para un potencial bien diseñado, la energía de la conformación nativa debe ser más favorable que para todas las quimeras. Estos métodos son interesantes porque permiten comparar la conformación nativa que corresponde a una secuencia con muchos otros plegamientos alternativos. Sin embargo, con este tipo de técnica no se puede comprobar la calidad del potencial como se hace en experimentos de minimización, donde un gran número de quimeras se generan automáticamente durante el muestreo conformacional. Esta generación de quimeras durante el propio proceso de minimización permite una búsqueda mucho más extensa en el espacio conformacional de la proteína. Normalmente, para ello se utilizan métodos de simulación tradicionales como la dinámica molecular o el método de Monte Carlo.

El objetivo de esta Tesis Doctoral es evaluar potenciales de plegamiento de proteínas basados en estructuras para los principales tipos de interacción: interacciones hidrófobas y enlaces de hidrógeno. Estas contribuciones a la energía global se estudian primero por separado, y a continuación, juntas. Así ofrecemos una comparación independiente de potenciales elaborados por distintos grupos de investigación. Para ello, en

vez de utilizar el método de dinámica molecular o el de Monte Carlo, usamos un método evolutivo —basado en algoritmos genéticos— diseñado por nosotros mismos que permite una evaluación sistemática de los potenciales. Los algoritmos genéticos permiten estudiar los potenciales de interacción realizando un muestreo que no depende de la temperatura, al contrario de lo que sucede con los métodos más habituales. Además, dada su rapidez y su capacidad de escapar de mínimos locales, son especialmente apropiados para el estudio del plegamiento.

En la primera parte de la Tesis (Capítulo 2) se describe el fundamento y el desarrollo del método computacional. Se trata de un método de tipo evolutivo que permite obtener la conformación de menor energía para un modelo de proteína y un potencial de interacción dado. En este método se utiliza una representación reducida de la geometría de la proteína, especialmente apropiada para evaluar las distintas contribuciones energéticas. La puesta a punto de la metodología supuso un largo periodo de investigación, en el que se probaron diferentes variantes del algoritmo. En esta memoria se incluye sólo un resumen de todo ese trabajo.

En el Capítulo 3 se incluye la puesta a punto de este método con un potencial de tipo Gō[98]. El tipo de potencial utilizado es especialmente apropiado para probar la eficiencia de un método de búsqueda conformacional[99]. Al contrario de lo que sucede con los potenciales estadísticos, tiene su origen en la conformación nativa de cada proteína considerada, por lo que no tiene ningún carácter predictivo. A la hora de interpretar los resultados, este tipo de potencial permite desvincular la eficiencia de la técnica de búsqueda y la capacidad del potencial de definir su mínimo en la conformación nativa. Así, podemos evaluar la eficiencia del algoritmo en cada una de las versiones descritas en el Capítulo 2.

Una vez puesta a punto la metodología con el potencial de tipo Gō, la aplicamos a

un grupo de potenciales basados en estructura para las interacciones entre cadenas laterales[92,96,100] Estos potenciales tratan de explicar el colapso hidrófobo, que es considerado la fuerza impulsora del proceso de plegamiento[24]. Esta parte del estudio se incluye en el Capítulo 4[101].

A continuación, nuestra investigación se ha centrado en la interacción de tipo enlace de hidrógeno, responsable, como hemos comentado, de la formación de los dos tipos principales de estructura secundaria[1]. En este caso no se estudiaron potenciales estadísticos, mucho más escasos en la bibliografía que para la interacción entre cadenas laterales. En cambio, seleccionamos una serie de modelos reducidos con distinto nivel de resolución[102–104]. Estos modelos pueden considerarse también basados en estructura en la medida en que parten de la observación de las características de las estructuras estabilizadas por los enlaces de hidrógeno en proteínas. Recogemos esta parte de nuestra investigación en el Capítulo 5[105].

Una conclusión extraída de las evaluaciones anteriores, la de potenciales hidrófobos y la de potenciales de enlace de hidrógeno, es el mejor comportamiento de unos potenciales sobre otros. En la última parte de la Tesis, que conforma el Capítulo 6, se incluyen los resultados de la evaluación conjunta del mejor potencial de campo medio y los mejores modelos para la interacción de puente de hidrógeno.

Esta memoria concluye con un Capítulo en el que se recogen las principales conclusiones de nuestro estudio.

Capítulo 2

Materiales y métodos

Como hemos comentado en el Capítulo 1, el objetivo de este estudio es evaluar potenciales de interacción para el estudio del plegamiento de proteínas. La condición que nos permite conocer la calidad de los potenciales para los distintos tipos de interacción es su capacidad de definir correctamente una superficie de energía. Como hemos dicho en la Introducción, Anfinsen, en su hipótesis termodinámica, establecía que el estado nativo de una proteína corresponde al mínimo de energía libre del sistema[4]. Por lo tanto, la condición mínima que debe satisfacer un potencial para una proteína dada es que tenga su mínimo energético en la conformación nativa. Partimos de esta premisa para desarrollar una herramienta de muestreo conformacional para localizar el mínimo energético con una serie de potenciales de interacción.

Otro aspecto sobre el que hemos hecho hincapié en el Capítulo 1 es la complejidad del espacio conformacional de las proteínas. Debido al elevadísimo número de grados de libertad, un muestreo sobre todos ellos en busca del mínimo energético resulta inabordable. Por este motivo, llevamos a cabo el muestreo conformacional para alcanzar el mínimo congelando un gran número de grados de libertad de la proteína. Así, en nuestro método una proteína se considera como un conjunto de fragmentos rígidos. La minimi-

zación consiste en localizar el ensamblaje de menor energía de estos fragmentos. En la sección siguiente explicamos otros motivos que hacen que esta aproximación resulte especialmente apropiada para nuestro estudio.

Los métodos más habituales para el muestreo del espacio conformacional de proteínas son el método de Monte Carlo y la dinámica molecular. En la Sección 1.5 comentábamos los problemas de este tipo de métodos en el caso del plegamiento de proteínas. El método de dinámica molecular resulta demasiado costoso computacionalmente como para utilizarlo en un muestreo sistemático del espacio conformacional de un grupo razonablemente amplio de proteínas para varios potenciales de interacción. Por otra parte, en el método de dinámica molecular hace falta resolver las ecuaciones de Newton para el movimiento. Para ello son necesarias funciones de energía que sean derivables, condición que no cumplen muchos de los potenciales de interacción que vamos a evaluar y que describiremos más adelante. En el caso del método de Monte Carlo, el problema reside en su tendencia a que la búsqueda se quede atascada en mínimos locales, al menos en las versiones más sencillas del método. Esto impide que se lleve a cabo un muestreo conformacional extenso que permita alcanzar con alguna certeza el mínimo de energía del sistema. Como una de las posibles alternativas a los dos métodos tradicionales, en este estudio utilizamos algoritmos evolutivos.

Los algoritmos evolutivos —los más conocidos de los cuales son los algoritmos genéticos— están basados en los mecanismos de la evolución natural que permiten la mejora de las especies[64]. Del mismo modo que en la Naturaleza la evolución se produce sobre especies, no sobre individuos, en estos algoritmos se maneja una población de posibles soluciones para el sistema. En la Naturaleza, la información relevante para la evolución está codificada en cromosomas, que portan la información genética. En los algoritmos evolutivos las soluciones del problema de búsqueda se codifican en cadenas de variables, que por analogía se llaman cromosomas. En nuestro caso, una solución

es una conformación de una proteína. Por tanto, en cada cromosoma de la población está codificada una conformación, en forma de una cadena de variables que la describe completamente.

Otro aspecto en que los algoritmos genéticos tratan de imitar la evolución natural es el transcurso del tiempo. En efecto, en la Naturaleza la evolución no se produce instantáneamente sino a lo largo de un periodo de tiempo. Análogamente, en los algoritmos evolutivos, aunque no existe el tiempo propiamente dicho, la evolución tiene lugar a lo largo de un cierto número de generaciones. Esta evolución se produce debido a que la información de los cromosomas de la población es modificada mediante la aplicación de unos operadores. Los operadores genéticos se denominan replicación, "crossover" o entrecruzamiento y mutación, e imitan una serie de mecanismos de la evolución natural. Estos operadores combinan y modifican las soluciones en una etapa de la optimización para generar nuevos individuos, como en la Naturaleza, mejor adaptados a su entorno. En el algoritmo, esta adaptación se cifra en el valor que adoptan los individuos de la población para una función, denominada función "fitness" o función de mérito, que es la magnitud que el algoritmo pretende optimizar. En nuestro caso, la elección obvia para la función de mérito es la energía de cada conformación. La aplicación de los operadores y la selección de los mejores individuos a lo largo de las generaciones permite que se alcance, eventualmente, la solución óptima para el problema de búsqueda.

Este Capítulo lo dedicamos al desarrollo y la completa descripción de la metodología que vamos a usar. En primer lugar describimos la aproximación utilizada para la búsqueda conformacional, el empaquetamiento de fragmentos, y el tratamiento simplificado de la geometría de la proteína. A continuación, tratamos el problema de la codificación, es decir, de representar conformaciones de la proteína como cadenas de variables del algoritmo genético. Finalmente, describimos punto por punto el funcionamiento de nuestras implementaciones del algoritmo. En este Capítulo damos cuenta

de las diferentes versiones del método desarrolladas en esta Tesis, desde el algoritmo genético inicial hasta la estrategia evolutiva.

2.1. Representación de la proteína

En el Capítulo 1 hemos hablado de los distintos niveles de resolución que pueden utilizarse para representar la estructura de una proteína. En general, el muestreo del espacio conformacional está muy vinculado al tipo de representación que se utilice para la cadena polipeptídica. En los estudios atomísticos se puede llevar a cabo un muestreo de un gran número de grados de libertad de la proteína y del disolvente. Debido a las exigencias computacionales de estos métodos, muchas veces se utilizan modelos más simplificados, en los que se realiza un muestreo sobre los ángulos de torsión de la cadena polipeptídica, o sobre una serie de ángulos y distancias virtuales entre centros de interacción del modelo, que agrupan grados de libertad reales de la proteína. Cada tipo de representación del espacio de búsqueda tiene sentido dentro de un modelo coherente para las interacciones. Por otra parte, y como veíamos también en la Introducción, siempre un mayor nivel de detalle trae consigo un mayor gasto computacional.

Desarrollamos nuestro método con el objetivo de evaluar modelos de interacción para proteínas. Con él queremos estudiar los distintos tipos de interacción independientemente: en primer lugar, las interacciones hidrófobas, y a continuación, los enlaces de hidrógeno. Nuestra intención es realizar un muestreo rápido sobre un número muy elevado de conformaciones para un conjunto razonablemente grande de proteínas. Por todo ello, resulta conveniente congelar un gran número de grados de libertad de la proteína, de tal manera que su geometría quede reducida a un determinado número de fragmentos rígidos. Así, el problema de búsqueda conformacional consiste en un muestreo sobre los posibles empaquetamientos de los fragmentos rígidos. Este método de ensamblaje de

fragmentos ha sido utilizado previamente por muchos otros autores, especialmente en predicción de estructura de proteínas[100,106–114]. Lógicamente, la simplificación del espacio conformacional de la proteína introducida condiciona el análisis de los resultados, lo que hace necesario que el diseño del modelo sea muy cuidadoso.

En nuestra aproximación, uno de los aspectos en que hay que tener especial cuidado es el tipo de proteína con el que se ponen a prueba los distintos potenciales de interacción. Una vez seleccionada una proteína en particular, hay que dividirla en fragmentos de manera coherente. Los fragmentos resultantes de esta división deben poder empaquetarse con el modelo para las interacciones que vayamos a evaluar. Por ejemplo, para estudiar de forma aislada potenciales de interacción hidrófobos, podemos considerar proteínas con estructura de tipo "coiled-coil" o de hélice superarrollada, como la proteína con código PDB 1ij3 que mostramos en la Figura 2.1 (a). Estas estructuras están formadas por hélices α largas e independientes, que oligomerizan debido a interacciones fundamentalmente hidrófobas[115]. Por ello, las hélices superarrolladas son un tipo estructural muy apropiado para evaluar potenciales que tratan de reproducir este tipo de interacción. En la representación de la Figura 2.1 (a) mostramos en diferentes colores las tres hélices independientes de la proteína 1ij3 que consideraríamos en nuestra aproximación. En lugar de proteínas con hélices superarrolladas en su estructura, podemos considerar proteínas globulares en cuya conformación nativa encontremos hélices empaquetadas. En la Figura 2.1 (b) representamos una proteína de este tipo, con código PDB 1abv. También para proteínas como esta, las hélices α se mantienen unidas en su estructura nativa por interacciones fundamentalmente hidrófobas. Como en el caso anterior, en nuestro método los fragmentos rígidos serían las hélices que mostramos en distintos colores en la Figura. Finalmente, en el caso de los enlaces de hidrógeno, podemos utilizar láminas β que encontremos en la estructura terciaria de una proteína, como la de la Figura 2.1 (c). Las láminas β están constituidas por una serie de hebras que

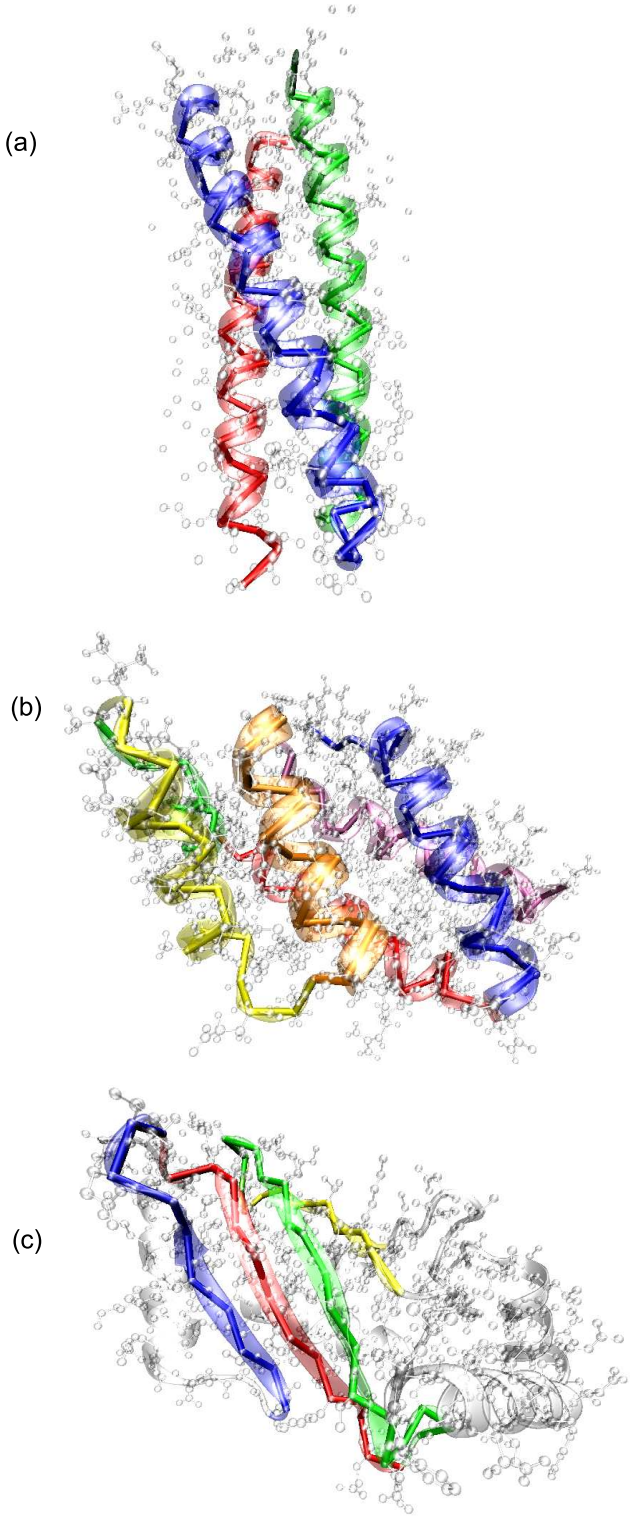


Figura 2.1: Tres proteínas que pueden utilizarse en el muestreo conformacional con el método de fragmentos rígidos. Para cada una, representamos todos sus átomos, sus elementos de estructura secundaria y la traza de carbonos-α de los residuos que integramos en los fragmentos congelados, que mostramos en diferentes colores. Códigos PDB: (a) 1ij3, (b) 1abv, (c) 1q34.

se unen entre sí por enlaces de hidrógeno. En este caso los fragmentos cuya geometría congelamos son las hebras β. Como hemos visto para estos tres ejemplos, en general en nuestro método vamos a realizar un muestreo conformacional limitado a los empaquetamientos de los elementos de estructura secundaria presentes en la conformación nativa de la proteína.

Otro aspecto que hay que definir en la representación de la proteína que usa el método es la resolución con que representamos los residuos de aminoácido. Según cómo estén definidas las interacciones en el potencial que vayamos a evaluar, serán unos u otros los centros que consideremos de cada residuo. Así, para todas las proteínas con las que trabajamos, tomamos la secuencia y las coordenadas de sus átomos directamente del Protein Data Bank (PDB)[56,57]. En unos casos, de toda esta información utilizaremos únicamente las coordenadas de los carbonos-α, en otros casos también las de los átomos de las cadenas laterales, mientras que en otros se tomarán todos los átomos del esqueleto de la proteína. Por tanto, el nivel de resolución de nuestro método varía en función del potencial que evaluemos.

2.2. Codificación del algoritmo genético

Como hemos comentado, el problema al que nos enfrentamos es localizar el empaquetamiento de fragmentos peptídicos de menor energía para una determinada proteína con un modelo para las interacciones. El método que utilizamos para llevar a cabo el muestreo conformacional es un método evolutivo, basado en algoritmos genéticos. En los algoritmos genéticos se maneja una población de cromosomas que representa un grupo de soluciones de un problema de búsqueda. Una codificación es la clave que permite transformar las posibles soluciones en una serie de números, escritos en sistema binario o como números reales, que la describan completamente. En un trabajo seminal sobre

algoritmos genéticos[64], Goldberg sugería que el único límite para diseñar nuevas codificaciones lo fijaba la imaginación de los programadores. Para nuestro algoritmo de minimización de la energía de proteínas, hemos desarrollado hasta tres codificaciones distintas con las que convertir en cromosomas las soluciones de nuestro problema de optimización, los empaquetamientos de fragmentos peptídicos. Todas ellas comparten unas características generales que describimos a continuación.

Para describir una conformación de la proteína con nuestro modelo de fragmentos rígidos, codificamos en los cromosomas dos grupos de tres variables por cada uno de los fragmentos, exceptuando uno de ellos, que permanece fijo con respecto al sistema de referencia. El primer conjunto de variables para un fragmento, (r, θ, φ), codifica las coordenadas polares esféricas de una posición del fragmento con respecto al sistema de referencia. Una vez determinada esa posición, el otro conjunto de variables, (Θ, Φ, Ψ), codifica los ángulos de Euler que permiten orientar el fragmento a partir de ese punto y con respecto a una orientación de referencia. Así, para una proteína que en nuestro método está dividida en N fragmentos, un cromosoma tiene $6\times(N-1)$ variables y se escribe como $\{r_1, \theta_1, \varphi_1, \Theta_1, \Phi_1, \Psi_1, r_2, \theta_2, \varphi_2, \Theta_2, \Phi_2, \Psi_2, ..., r_{N-1}, \theta_{N-1}, \varphi_{N-1}, \Theta_{N-1}, \Phi_{N-1}, \Psi_{N-1}\}$. A partir de estas características comunes, las distintas codificaciones funcionan como describimos a continuación.

2.2.1. Codificación externa

Hemos denominado "externa" a esta codificación porque el origen de coordenadas para disponer los fragmentos peptídicos en el espacio se encuentra en un punto que no coincide con un átomo de la proteína. El origen de coordenadas es, por tanto, externo a ella, y lo ubicamos, para mayor simplicidad, en el centro geométrico de la proteína en su conformación nativa.

Con respecto a este sistema de referencia, el primer fragmento se mantiene fijo.

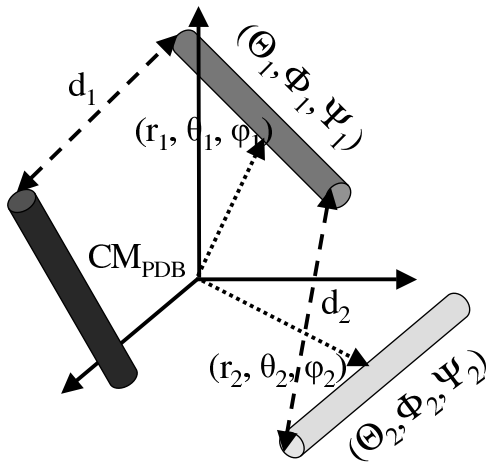


Figura 2.2: Representación esquemática de la definición de las variables con la codificación externa.

Esto significa que su disposición en el espacio es la misma que dan las coordenadas del PDB. Colocamos el resto de fragmentos usando el conjunto de variables correspondiente. Un primer grupo de variables, $(r_i, \theta_i, \varphi_i)$, describe las coordenadas esféricas del centro del fragmento $(i+1)$ respecto al centro geométrico de la proteína, como se indica en la Figura 2.2. El ejemplo corresponde a una proteína de la que se consideran tres fragmentos rígidos, que se representan en la Figura como cilindros. El segundo grupo de variables de este fragmento, $(\Theta_i, \Phi_i, \Psi_i)$, corresponde a los ángulos de Euler que permiten establecer su orientación. De acuerdo con los valores de estos ángulos, cada fragmento de la proteína rota con respecto a su orientación original. Esta orientación original es la que tiene el fragmento en la conformación nativa tomada del PDB. Esto no determina en absoluto el resultado de la búsqueda porque la conformación nativa —que correspondería a (Θ, Φ, Ψ)=(0,0,0)— no se utiliza como uno de los individuos de la población inicial.

Como hemos comentado, en las tres primeras variables de cada grupo de seis está codificada la posición del centro de un fragmento. La distancia con respecto al centro

geométrico de la proteína nos la da la primera de las variables del grupo, r. Como vamos a buscar empaquetamientos que minimicen la energía, no nos interesa que se generen conformaciones muy poco compactas. Para evitar que se pierda tiempo de la minimización con estas conformaciones, acotamos el valor de r. Calculamos la distancia de cada fragmento al centro geométrico de la proteína en la conformación nativa, se determina cuál es la mayor de todas ellas, y el valor máximo de r se fija en cinco veces su valor. Por otra parte, el dominio de los ángulos θ y φ es el habitual: $[0,\pi]$ y $[0,2\pi]$, respectivamente. Esto ha de dar suficiente libertad a los fragmentos para que se puedan generar todo tipo de conformaciones con respecto al sistema de referencia.

Con esta codificación, cada uno de los fragmentos peptídicos de una proteína puede moverse con gran autonomía con respecto a los demás. Esto es químicamente aceptable cuando la proteína que estudiamos está formada por cadenas independientes, por ejemplo en el caso de las hélices superarrolladas que representábamos en la Figura 2.1 (a). Por tratarse de hélices independientes, no importa que no haya una restricción en el desplazamiento de cada fragmento. Por el contrario, en el caso de que los fragmentos sean sucesivos en la secuencia de una sola cadena, como en el caso de la proteína globular de la Figura 2.1 (b), esta codificación podría originar problemas. Los fragmentos podrían alejarse entre sí hasta el punto de que la distancia entre extremos supusiese la ruptura de la cadena. Por eso, en estos casos hay que imponer alguna restricción a la disposición relativa de los fragmentos peptídicos. Así, la búsqueda se concentrará en aquellas estructuras que respeten la conectividad de la cadena. En nuestro algoritmo, esta restricción se introduce en el cálculo de la energía para cada par de fragmentos sucesivos en la cadena. En el caso de que la distancia d_i entre el último residuo del fragmento i y el primero del $(i+1)$ supere la máxima longitud que puede tener el lazo que los une en una conformación extendida, se penaliza la conformación correspondiente. Si la distancia d_i es más corta de lo admisible, también hay que introducir algún tipo de

penalización. Finalmente, si la proteína con la que trabajamos tiene fragmentos separados por un pequeño número de aminoácidos, muchas conformaciones de la proteína pueden penalizarse por no respetar la conectividad entre fragmentos. En tal caso, es más conveniente utilizar otra codificación en el algoritmo.

2.2.2. Codificación interna simple

Como en la codificación externa, en la codificación interna simple los residuos del primer fragmento quedan fijos en la posición en que aparecen en el archivo del PDB. En este caso, para colocar cualquier fragmento peptídico, el origen de coordenadas se sitúa sobre el fragmento anterior, y así se van colocando sucesivamente. Precisamente porque se pivota sobre un átomo del fragmento anterior para disponer cada fragmento en el espacio, hemos denominado a esta codificación "interna".

El grupo de variables $(r_i, \theta_i, \varphi_i)$ se utiliza para colocar el primer residuo del fragmento $(i+1)$ partiendo de la posición del último residuo del fragmento i. En esta codificación, el valor de la primera de las variables de este grupo, r_i, está acotada como lo estaba d_i en la codificación externa. Así, el valor mínimo de la distancia r_i es la distancia entre dos carbonos-α de residuos contiguos que forman un enlace peptídico "trans", que es de 3.8 Å; el valor máximo, la distancia de la conformación extendida de todos los enlaces que haya en el lazo que une los fragmentos. Al contrario de lo que sucedía en la codificación interna, las restricciones debidas a la conectividad de la cadena están implícitas en la codificación. Después, el segundo grupo de tres variables $(\Theta_i, \Phi_i, \Psi_i)$ permite establecer la orientación del fragmento $(i+1)$ con respecto a la original. Mostramos el significado de las variables en esta codificación para una proteína de tres fragmentos en la Figura 2.3 (a). Excepto en el caso de la variable r, que como ya hemos mencionado es dependiente del número de residuos entre fragmentos, las variables tienen el mismo dominio que hemos definido para ellas con la codificación externa.

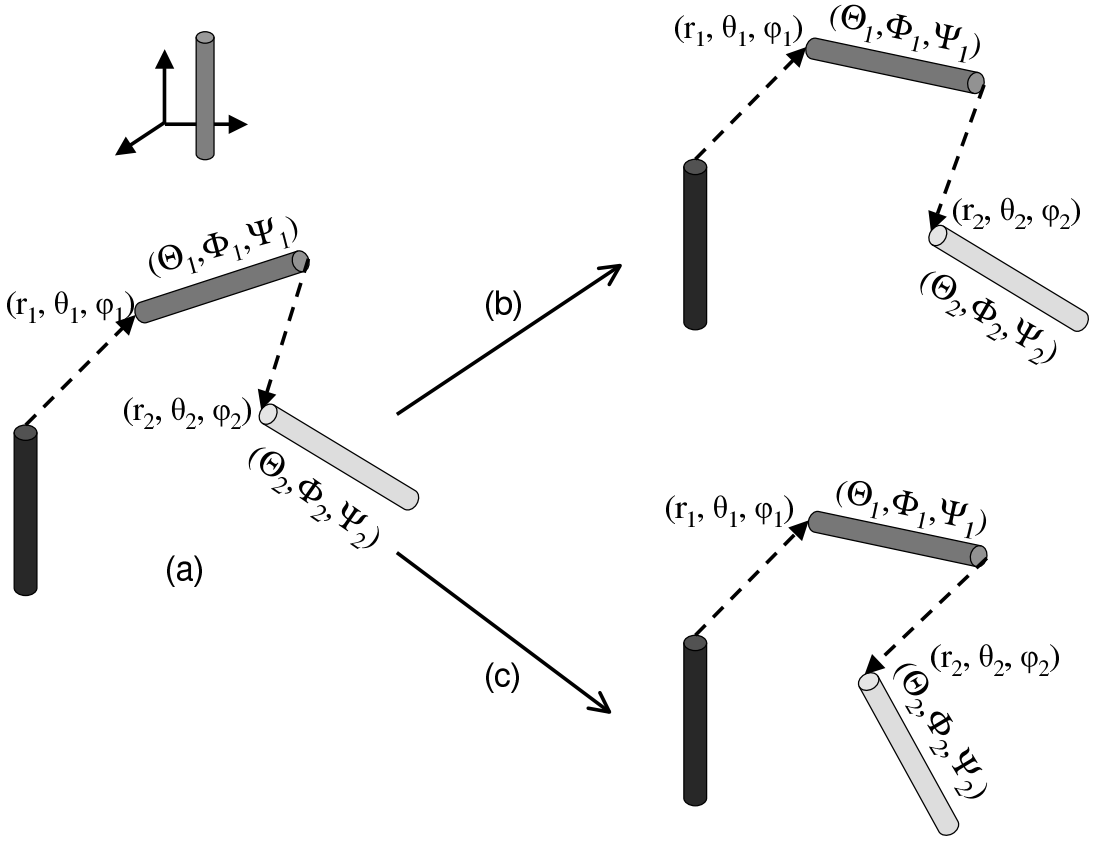


Figura 2.3: Representación de las dos variantes de la codificación interna para una proteína de tres fragmentos. (a) Significado de las variables. (b) y (c) Efectos que un cambio en los ángulos de Euler del fragmento 2 tendrían para la variante simple y compleja, respectivamente (ver texto).

En el diagrama de la Figura 2.3 (b) hemos tratado de representar lo que sucedería ante un cambio en la cadena de variables. El cambio que representamos corresponde al primer grupo de ángulos de Euler, $(\Theta_1, \Phi_1, \Psi_1)$. Estas tres variables describen la orientación del segundo fragmento peptídico, que rota conforme a los nuevos valores. Al modificarse esta orientación, también el tercer fragmento se desplaza, pero conserva la posición de su extremo inicial con respecto al extremo terminal del fragmento anterior, y su orientación relativa al sistema de referencia externo, como se indica en la Figura 2.3 (b).

Hemos estudiado cómo esta codificación permite que se generen nuevas topologías a lo largo de sucesivas generaciones del algoritmo genético, en las que cambios en los cromosomas como el descrito en el párrafo anterior pueden ser muy habituales. Pero también pueden darse situaciones en las que resulte difícil empaquetar varios elementos de estructura secundaria a partir de una conformación que ya tuviese varios fragmentos interaccionando. Supongamos que, en la conformación que consideramos, los fragmentos 2 y 3 estuviesen próximos. Al cambiar la orientación del fragmento 2, perderíamos un buen empaquetamiento, ya que la orientación de 3 se fija con respecto a un sistema de referencia externo. Debido a este inconveniente creamos una nueva codificación.

2.2.3. Codificación interna compleja

Esta codificación trata de subsanar los problemas de la interna simple. Las variables de un cromosoma tienen el mismo significado que en la anterior. Sin embargo, es diferente la relación entre las variables (Θ,Φ,Ψ) que determinan la orientación de los fragmentos. Pensemos en el caso que mostrábamos antes, para una conformación de una proteína en la que se modifica uno de los ángulos de uno de los fragmentos. En esta codificación todos los fragmentos subsiguientes a aquel cuyos ángulos varían rotan coordinadamente con él. Esto lo hemos tratado de representar para el caso de tres fragmentos en la Figura 2.3 (c).

Supongamos que en la conformación de partida se modifica uno de los ángulos de Euler del segundo fragmento, como veíamos en el apartado anterior. Esto supone una rotación del fragmento 2. En esta codificación, en efecto se produce un giro del fragmento 2, pero además gira con él el fragmento 3, manteniendo la misma orientación relativa con respecto al 2 que tenía antes de la mutación. De este modo pretendemos que la aplicación de los operadores no resulte destructiva para los dominios con interacciones favorables de una solución. Los empaquetamientos correctos que se hayan generado en cada caso quedan así salvaguardados.

Con la codificación interna compleja intentamos subsanar los problemas de la codificación interna simple. Por tratarse de dos codificaciones complementarias, las utilizamos conjuntamente. Esto significa que cada uno de los cromosomas de la población tiene dos significados, es decir, corresponde a dos conformaciones diferentes de la proteína. Se calcula la energía para ambas y consideramos que la estructura asignada a ese individuo de la población es la de la conformación de menor energía.

2.3. Algoritmo genético

Acabamos de describir las distintas codificaciones que permiten transformar conformaciones de la proteína en cromosomas, es decir, en cadenas de variables. Esto permite que abordemos nuestro problema de muestreo conformacional utilizando un algoritmo genético. Como hemos comentado anteriormente, en un algoritmo genético se utiliza una población de cromosomas, es decir, se maneja un cierto número de cadenas de variables. Para que se produzca la evolución de esta población hacia el óptimo, sobre los cromosomas actúa una serie de operadores. Los operadores genéticos, llamados replicación, "crossover" o entrecruzamiento, y mutación, transforman las cadenas de variables de distintas maneras. Así, se modifican los cromosomas de una población de partida

y se obtiene una progenie de nuevos individuos. Para que el sistema evolucione hacia su óptimo, ha de favorecerse que los individuos más aptos prosperen y que los menos aptos vayan desapareciendo. En el algoritmo, esto se consigue seleccionando los mejores individuos entre la población de partida y la progenie generada mediante la aplicación de los operadores. Los mejores individuos, en este caso, son los que tienen un valor más próximo al óptimo para la función "fitness" o función de mérito. En nuestro caso, la función de mérito es lógicamente la energía. A lo largo de las generaciones, la aplicación de los operadores y la selección de individuos más aptos permite que nos vayamos acercando al mínimo energético de un determinado potencial.

En el algoritmo que hemos desarrollado, en primer lugar generamos una población inicial de n_c cromosomas, es decir, n_c cadenas de variables. Para una proteína que dividimos en N fragmentos, cada una de las cadenas está formada por $6 \times (N-1)$ variables que generamos al azar entre 0 y 1. Estos valores entre 0 y 1 se cambian de escala para obtener los valores de distancias y ángulos en los intervalos correspondientes. A partir de los valores para las variables, obtenemos la conformación correspondiente a cada cadena conforme a la codificación que utilicemos. Para cada conformación de la población inicial calculamos la energía según el potencial, que es su valor para la función de mérito. Como hemos comentado, la población inicial se genera al azar. Por tanto, en esta población puede haber tanto conformaciones de alta como de baja energía. Las soluciones de la población inicial comienzan a transformarse mediante la aplicación de los operadores genéticos, cuyo funcionamiento describimos a continuación. En la Figura 2.4 mostramos una representación esquemática del funcionamiento de los operadores genéticos sobre dos cromosomas en nuestro método.

1. Replicación: Como mostramos en la Figura para el ejemplo con una población de dos cromosomas, este operador consiste en copiar directamente individuos de la población de partida en la progenie. En nuestro caso se copian a la progenie

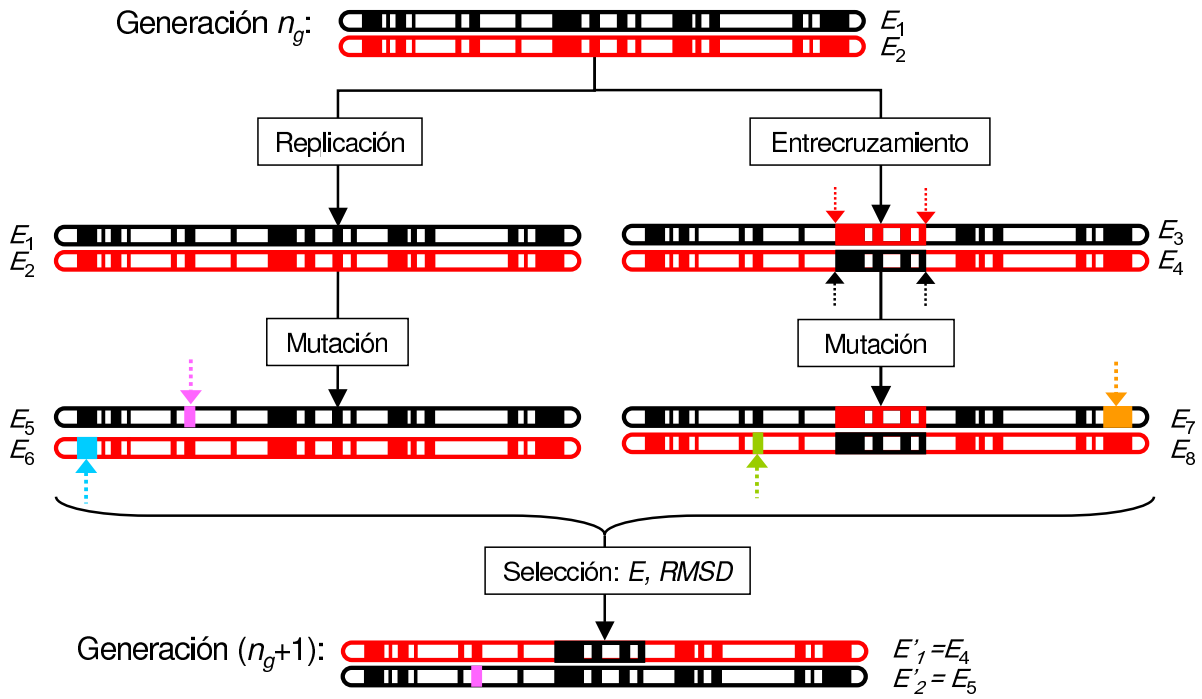


Figura 2.4: Funcionamiento de los operadores en el algoritmo genético, para una población de n_c=2 individuos. Cada individuo viene representado por un cromosoma de distinto color, cuya energía es E_i. Aplicando los operadores genéticos sobre la población en la generación n_g se obtiene una progenie de 4×2 cromosomas. Las flechas indican los puntos de la cadena en los que actúan los operadores. De la progenie se seleccionan 2 individuos para formar la nueva población en la generación $(n_g + 1)$, con un criterio combinado de energía y de estructura (ver texto).

los n_c individuos de la población. Así, nos aseguramos de que en principio no se pierda ninguna información que pueda resultar valiosa por la aplicación de los otros operadores.

2. Entrecruzamiento: Este operador permite el intercambio de información entre pares de individuos. En la Figura 2.4 mostramos con un código de colores cómo se intercambian fragmentos de cromosoma tomados de la población de partida. En el algoritmo, los fragmentos entrecruzados contienen un cierto número de variables. Con este operador generamos otros n_c individuos. En general, nos interesa que compartan la información de sus cadenas los mejores individuos de la población. Así, para seleccionar cada individuo se utiliza un mecanismo de *rueda de ruleta*[64]. Este mecanismo permite seleccionar con mayor probabilidad los individuos con mejor valor para la función de mérito. Cada par de cromosomas seleccionados se corta en un punto, seleccionado al azar, y los fragmentos entre ese punto y el final del cromosoma se intercambian. Como alternativa, indicada en el ejemplo de la Figura 2.4, las cadenas también pueden cortarse en dos puntos e intercambiar un fragmento de menor tamaño de la cadena. En nuestro algoritmo este fragmento es de una sola variable. Los dos tipos de entrecruzamiento tienen efectos diferentes. El de un punto permite el intercambio de fragmentos grandes, con varias variables, y puede tener mucha importancia en los primeros estadios de la optimización. El de dos puntos, al intercambiar fragmentos pequeños, favorece el ajuste fino de las soluciones de la población. Un parámetro ajustable permite decidir qué tipo de entrecruzamiento se realiza: con una probabilidad p_{cross} se lleva a cabo el de dos puntos, y con una probabilidad $(1 - p_{cross})$, el de un punto.

3. Mutación: Con el operador de mutación se introduce nueva información en las cadenas, provocando cambios en alguna de las variables del cromosoma mutado.

Se toman copias de los individuos procedentes de los pasos 1 y 2, y se modifican sus variables con una cierta probabilidad p_{mut} pequeña. En la Figura 2.4 representamos las mutaciones como cambios de color en los cromosomas. En nuestro algoritmo, generamos así $2n_c$ individuos, que no remplazan a los de los pasos 1 y 2 de los que proceden. Así, preservamos los individuos de la población de partida y los obtenidos por entrecruzamiento. En nuestro método, el operador de mutación actúa sobre un cromosoma remplazando el valor de una variable por otro número elegido al azar en el mismo intervalo. Dado que realizamos cambios en variables con una probabilidad pequeña, podríamos dejar inalteradas muchas cadenas. Esto supondría una replicación encubierta de individuos. Evitamos este efecto garantizando que se mute al menos una posición de cada cadena.

Aplicando estos operadores sobre la población se obtiene una progenie formada por un total de $4n_c$ cromosomas, muchos más que la población inicial. Para evitar el crecimiento descontrolado de la población, se lleva a cabo una selección de individuos. En general, esta selección se realiza conforme a su clasificación de acuerdo con su valor para la función de mérito. De esta manera, seleccionaríamos n_c individuos siguiendo el orden de energía, de menor a mayor. Hemos observado que esto conduce a una dramática disminución en la variabilidad de la población. En nuestro caso, la población se va enriqueciendo en muchas conformaciones iguales o muy parecidas de la proteína. Con frecuencia esto conduce a la convergencia de la población hacia un mínimo local para la energía. Para tratar de evitar problemas de convergencia prematura, combinamos el orden de selección por energía con otro criterio que garantice la diversidad de la población. Hemos escogido una función de diferencia entre conformaciones, la desviación cuadrática media de las posiciones de los carbonos-α ($RMSD$). Esta magnitud refleja cambios globales en la estructura, y se utiliza muy habitualmente en el análisis conformacional de proteínas. Para obtener el valor de $RMSD$ entre dos conformaciones 1 y 2

de una proteína con n residuos, en primer lugar se busca la mejor superposición de sus carbonos-α[116]. Después, la desviación cuadrática media se calcula como

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{1,i} - r_{2,i})^2}, \tag{2.1}$$

donde $r_{k,i}$ es la posición del carbono-α del residuo i en la conformación k. Al añadir un criterio estructural al criterio energético, la selección queda como sigue. En primer lugar seleccionamos el individuo de menor energía. Para cada uno de los siguientes en orden creciente de energía, comprobamos que su $RMSD$ con respecto a todos los individuos ya seleccionados es mayor que un determinado umbral. Si cumple esta condición, entra a formar parte de la población; si no, se descarta. Esto lo hacemos hasta que hayamos seleccionado una nueva población de n_c individuos. En el caso de que no se pueda cumplir este criterio estructural para ninguna de las conformaciones que queden por seleccionar, se selecciona sólo por el criterio energético. En el ejemplo de la Figura 2.4, el primero de los individuos seleccionados es por tanto el de menor energía, E_4, procendente del entrecruzamiento. Para seleccionar el siguiente, se sigue el orden de energías, pero se calcularía el valor de $RMSD$ con respecto al individuo seleccionado.

Este procedimiento se repite iterativamente hasta que se completa un número de generaciones. Mediante esta repetición del proceso —replicación de soluciones, entrecruzamiento, mutación, selección— la población de cromosomas va mejorando por el efecto de los operadores. La mejora de la población se traduce en el acercamiento de alguno de sus individuos al óptimo de la función de mérito. Esto supone que los empaquetamientos de los fragmentos de la proteína cada vez se encuentran más próximos al mínimo energético para un potencial de interacción dado.

2.4. Funcionamiento del algoritmo genético

Hemos implementado el algoritmo descrito en un programa informático. Con él, hemos llevado a cabo la minimización de la energía para una proteína, la subunidad δ de la ATP sintasa F1F0 de *E coli*, con código PDB 1abv, que mostrábamos en la Figura 2.1 (b). En la estructura terciaria de esta proteína encontramos seis hélices α. En nuestro modelo, por tanto, la proteína consta de 6 fragmentos. En este caso hemos utilizado como función de mérito para el método un potencial que nos permite poner a prueba su eficiencia. Se trata de un potencial de tipo Gō[98], que describiremos en profundidad en el siguiente Capítulo. Por construcción, este tipo de potencial tiene su mínimo en la conformación nativa de la proteína, y el valor de la función de mérito en el mínimo es 0.

Algunos parámetros del algoritmo genético los hemos fijado de acuerdo con resultados obtenidos en cálculos preliminares: el tamaño de la población se ha fijado en n_c=100 individuos, y el umbral de $RMSD$ en la selección de conformaciones para obtener una nueva generación, en $RMSD_{ij}^{min}$=1 Å. Estos valores garantizan una notable variabilidad en la población, sin que su tamaño crezca mucho, lo que supondría un aumento en el tiempo invertido en el cálculo. Hemos obtenido resultados para diferentes valores de otros parámetros, como la probabilidad de mutación, p_{mut}, y las probabilidades de entrecruzamiento de dos puntos, p_{cross}, y de un punto, $(1 - p_{cross})$. Esto nos permite conocer cómo depende la eficiencia del algoritmo de estos parámetros. La optimización se ha realizado cinco veces para cada conjunto de parámetros. Las cinco ejecuciones son iguales excepto en el número semilla que necesita el generador de números pseudo-aleatorios. Este número semilla condiciona todas las operaciones estocásticas que se realizan a lo largo de una ejecución del programa, entre ellas la producción de una población inicial de cromosomas.

Mostramos los resultados de estos cálculos en la Figura 2.5. En cada representación

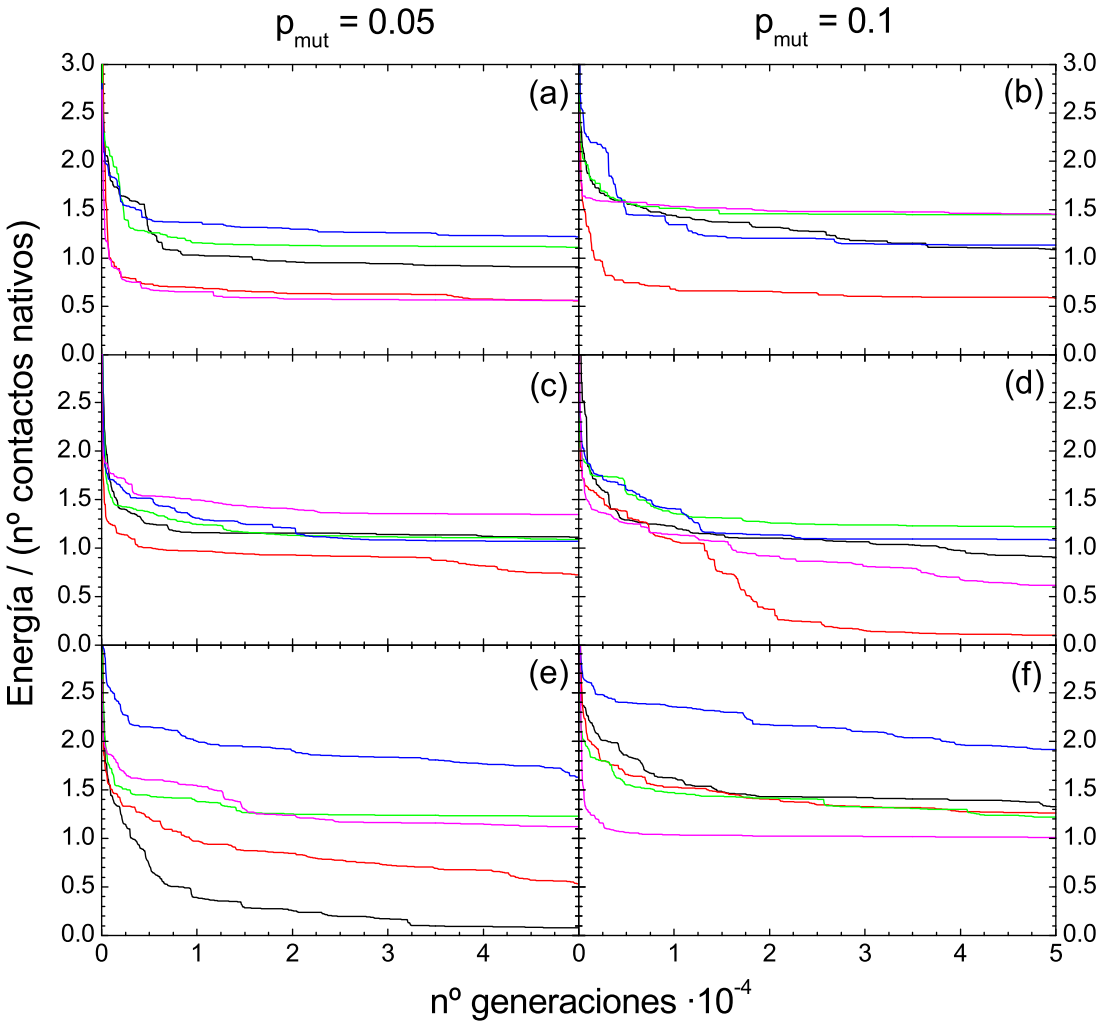


Figura 2.5: Valores de la energía mínima normalizada en cada generación a lo largo de la minimización con el algoritmo genético para 1abv. En cada caso se representan resultados de cinco ejecuciones diferentes. Los distintos paneles corresponden a distintos valores de la probabilidad de mutación p_{mut} (en la parte superior de la Figura) y la probabilidad de entrecruzamiento de dos puntos: (a) y (b) p_{cross}=0; (c) y (d) p_{cross}=0.5; (e) y (f) p_{cross}=1.

mostramos el valor de energía de la mejor conformación de la población a lo largo de las generaciones para las cinco minimizaciones que se han llevado a cabo para cada conjunto de parámetros. En algunas ejecuciones se alcanzan mínimos energéticos muy bajos, como en el caso de p_{mut}=0.1 y p_{cross}=0.5 (ver Figura 2.5 (d)), o el de p_{mut}=0.05 y p_{cross}=1.0 (ver Figura 2.5 (e)). Sin embargo, en la mayoría de las optimizaciones los valores finales de la energía normalizada a los que ha llegado el algoritmo son superiores a 1. Esto supone que la búsqueda se queda estancada en un mínimo local. Además, para cualquier conjunto de parámetros, cada vez que ejecutamos el algoritmo encontramos resultados distintos. Esta falta de convergencia supone que los resultados del algoritmo genético dependen de manera muy importante del azar.

Los resultados podrían mejorarse aumentando el tamaño de la población, pero esto haría que se disparase el tiempo de cálculo. También podría incrementarse el número de generaciones preestablecido. Pero a pesar de todo ello seguiríamos sin tener garantía de que el valor alcanzado correspondiese al mínimo energético global para la proteína.

2.5. La estrategia evolutiva

En los resultados del algoritmo genético, hemos visto que las distintas minimizaciones realizadas con un mismo conjunto de parámetros casi nunca convergen hacia un mismo valor de la energía. Cada una de las minimizaciones corresponde a una exploración distinta del espacio conformacional. El muestreo podría enriquecerse si esas exploraciones independientes compartiesen la información de manera conveniente. Hemos elaborado un método que se asemeja al "modelo de islas"[117]. En este modelo, varias subpoblaciones del sistema son estudiadas en paralelo, y de cuando en cuando se produce la migración de un individuo de una subpoblación a otra. Utilizando esta idea, hemos desarrollado un nuevo método cuyo funcionamiento esquematizamos en la Figura 2.6.

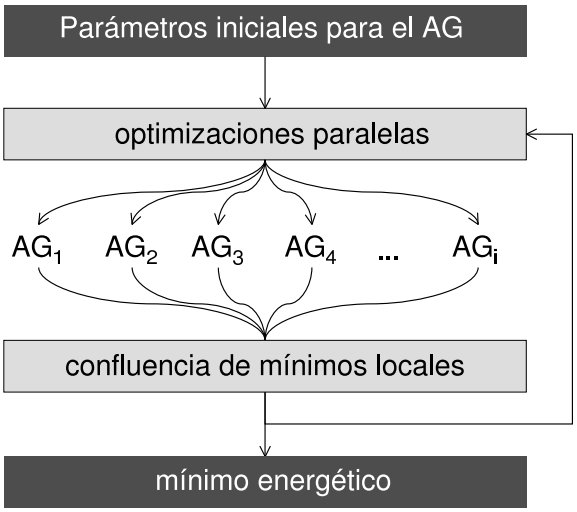


Figura 2.6: Diagrama de flujo de la estrategia evolutiva. AG_i representa a las optimizaciones independientes del algoritmo genético.

En nuestro nuevo método, un conjunto de parámetros común alimenta varias ejecuciones independientes del algoritmo genético. El conjunto de parámetros difiere de una optimización a otra únicamente en el número semilla. Recordemos que el número semilla determina las operaciones estocásticas, entre ellas la generación de una población inicial. Por tanto, los cromosomas de esta población son diferentes para cada una de las optimizaciones independientes. Como hemos descrito en la sección anterior, cada una de las ejecuciones del algoritmo realiza una optimización en la que progresivamente la búsqueda se va concentrando en un cierto tipo de conformación de la proteína. Previsiblemente, al cabo de n_{gener} generaciones, estas minimizaciones se quedarán atascadas en mínimos locales como sucedía con el algoritmo genético sencillo. En la estrategia evolutiva, los i mínimos resultantes de las optimizaciones independientes se reúnen para formar una población intermedia. Esta población intermedia de i individuos se utiliza para alimentar de nuevo las i minimizaciones, y cada nueva población inicial se completa, hasta alcanzar n_c, con individuos nuevos generados al azar en cada una de las minimizaciones independientes. Este mecanismo de confluencia se repite hasta que no

se observa una mejora apreciable en la energía mínima. Hemos establecido la mejora mínima en un 1 % con respecto al mínimo global del anterior ciclo de minimizaciones independientes.

2.6. Funcionamiento de la estrategia evolutiva

Ponemos a prueba la estrategia evolutiva con la misma función de mérito que hemos utilizado para comprobar la eficiencia del algoritmo genético sencillo. Como hemos dicho, se trata de un potencial de tipo Gō, muy apropiado para poner a prueba un método de muestreo por tener el mínimo definido en la conformación nativa con un valor conocido para la energía. Hemos realizado cálculos para la misma proteína de 6 fragmentos, 1abv, para un conjunto de valores de los parámetros fijo. Como en el caso del algoritmo genético sencillo, el tamaño de la población es n_c=100 individuos. Para mantener la diversidad de la población, hemos impuesto un valor relativamente alto para la probabilidad de mutación (p_{mut}=0.1). Así, este operador puede resultar un verdadero generador de nuevas líneas de búsqueda. Hemos permitido que los dos tipos de entrecruzamiento sean igualmente probables (p_{cross}=0.5), de manera que el algoritmo explore el mayor número de alternativas posible. Cada ejecución de la estrategia evolutiva consta de 10 minimizaciones independientes realizadas por el algoritmo genético. Estas minimizaciones independientes tienen una duración de 500 generaciones entre dos operaciones de confluencia. Hemos llevado a cabo el mismo experimento cinco veces.

En la Figura 2.7 presentamos los resultados de una de las optimizaciones con la estrategia evolutiva. Mostramos de nuevo la representación de la energía mínima a lo largo de la minimización, pero ahora para sólo una ejecución de la estrategia evolutiva. Se puede observar cómo después de cada 500 generaciones aparece un escalón en la representación de la energía. En ese momento, las diez ejecuciones independientes del

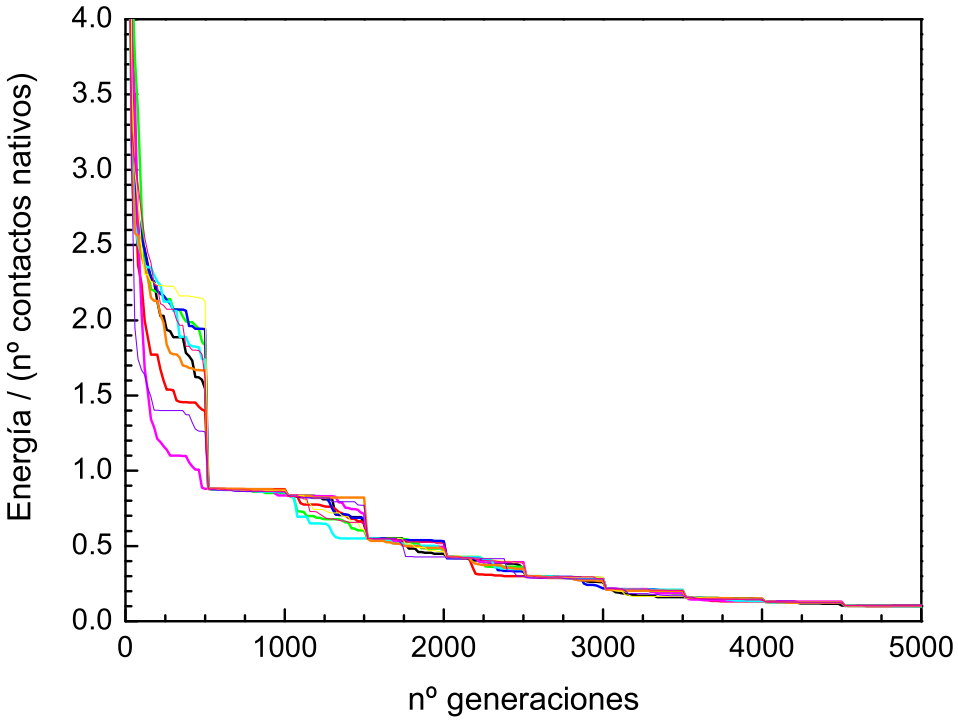


Figura 2.7: Valores de la energía mínima normalizada a lo largo de la evolución de una optimización realizada con la estrategia evolutiva para 1abv, una proteína de 6 fragmentos.

algoritmo genético ponen en común sus individuos más adaptados. Estos individuos pasan a integrar una población intermedia que alimenta de nuevo las minimizaciones independientes. En cada nueva minimización, la población intermedia formada por diez cromosomas se enriquece con noventa nuevos cromosomas generados aleatoriamente. Como las diez ejecuciones comparten los individuos que eran los mejores en cada una de ellas, el valor de energía de partida después de la confluencia es común. Tras otro ciclo de 500 generaciones, después de un total de 1000, se produce la siguiente confluencia, y así sucesivamente.

El valor final de la energía mínima es un orden de magnitud menor del obtenido en promedio en el caso del algoritmo genético. Para las otras cuatro minimizaciones que se han realizado con la estrategia evolutiva, el valor mínimo de energía también corresponde a una conformación con energía menor de la unidad. Los resultados son

razonablemente reproducibles y el funcionamiento del algoritmo es bastante rápido.

Queda probado así que con respecto a la situación que encontrábamos en el caso del algoritmo genético, hemos conseguido una mejora importante en la eficacia de la búsqueda con la estrategia evolutiva. Esta mejora se debe a la combinación de la operación de convergencia con la reconstrucción de la población, que permiten la diversificación del muestreo. Esta diversificación deriva del uso de los mejores individuos de las distintas minimizaciones realizadas por las ejecuciones independientes del algoritmo genético y un grupo muy amplio de cromosomas nuevos, que introducen información nueva en la población. Como en la evolución natural, en nuestro método evolutivo de minimización la variabilidad de la población es un factor clave para su buen rendimiento en espacios de búsqueda complicados. Sólo con un algoritmo que combina la exploración de la superficie de energía y la explotación de los mínimos locales hemos sido capaces de alcanzar la solución óptima del sistema.

Capítulo 3

Evaluación de la estrategia evolutiva con un potencial de Gō

En el Capítulo 2 hemos explicado detenidamente el método de minimización de la energía que hemos desarrollado para la evaluación de potenciales de plegamiento. Para conocer la confianza que nos ofrecen sus resultados, debemos poner a prueba nuestro método de muestreo conformacional. Este capítulo de la Tesis Doctoral lo dedicamos a la puesta a punto de nuestro método.

En nuestra estrategia evolutiva, el valor de la función "fitness" o función de mérito para una conformación es igual a su energía. Como ya hemos hecho para obtener los resultados del Capítulo anterior, aquí empleamos como función de mérito un potencial *ad hoc* de tipo Gō[98]. Este tipo de potencial está basado en la conformación nativa de la proteína, en la que tiene su mínimo energético. Obviamente, los potenciales de este tipo carecen de capacidad predictiva, pero generan una superficie de energía compleja, semejante a la que pueden engendrar los potenciales realistas que queremos evaluar con nuestro método. Por ello, este tipo de función constituye un test óptimo de la eficiencia del algoritmo.

También en el Capítulo anterior mostrábamos las tres codificaciones que hemos diseñado. Estas tres codificaciones nos permiten transformar conformaciones de la proteína en cromosomas que nuestro algoritmo evolutivo puede manejar. Para cualquiera de ellas, el significado de un cromosoma es el mismo, un empaquetamiento de fragmentos peptídicos. Sin embargo, las variables de los cromosomas que permiten obtener estos empaquetamientos tienen significados sutilmente diferentes en unas y otras. En este Capítulo estudiamos también cómo estas codificaciones de la estrategia evolutiva influyen en el rendimiento del algoritmo.

3.1. Función de mérito

Por tratarse de métodos de optimización, los algoritmos de la familia evolutiva requieren la definición de una función de mérito o función "fitness". En nuestro caso, como queremos explorar una superficie de energía para alcanzar su mínimo, la función de mérito es igual a la energía. A cada uno de los individuos de la población —a cada una de las conformaciones de la proteína— le corresponde un valor de la función de mérito. Nuestro objetivo es la evaluación de modelos de interacción realistas, por lo que podríamos aplicar directamente el método evolutivo a uno de ellos y minimizar la energía para un grupo de proteínas. Si lo hiciésemos, nos encontraríamos con que a la hora de analizar los resultados resulta difícil separar la eficacia del método de muestreo y las posibles imprecisiones en la definición de la superficie de energía. Por este motivo es necesario evaluar nuestro método con una función energética sencilla, cuyo mínimo es conocido.

El tipo de potencial que vamos a utilizar como función de mérito del algoritmo tiene su origen en los trabajos de Nobuhiro Gō *et al.*[98]. Se trata de un potencial *ad hoc*, es decir, basado en la estructura nativa de una proteína determinada y por tanto específico de ésta. Para este tipo de potenciales, las conformaciones de la proteína reciben

valores crecientes de energía en función de su diferencia con la nativa, de modo que, por construcción, el mínimo energético está en la conformación nativa. La diferencia de las conformaciones con respecto a la nativa viene cifrada en una serie de contactos, es decir, pares de centros de interacción que están en la conformación nativa a una distancia menor de un determinado valor. Para cada par ij de centros se define la distancia de equilibrio como la distancia que los separa en la conformación nativa (d_{ij}^{PDB}). Una función creciente determina cómo, al ir alejándose de esa situación de equilibrio, aumenta la energía asignada a ese par de centros. Los modelos de tipo Gō se han utilizado muy frecuentemente en la simulación del plegamiento de proteínas por su capacidad de definir un mínimo profundo y proporcionar un paisaje de energía ideal[118].

En nuestra implementación, para una determinada conformación de la proteína calculamos la energía como una suma de términos entre pares de centros de interacción, que en este caso son los carbonos-α. El valor de cada uno de estos términos es función de la diferencia entre las distancias entre pares de centros en la conformación que se está estudiando y esas mismas distancias en la estructura nativa. La expresión para el cálculo del término energético E_{ij}, correspondiente al par de centros de interacción ij, es:

$$E_{ij} = \omega_{ij}(d_{ij} - d_{ij}^{PDB})^2 \tag{3.1}$$

En esta ecuación, d_{ij} es la distancia entre carbonos-α en la conformación evaluada, y d_{ij}^{PDB} es la distancia a la que esos mismos centros se encuentran en la conformación nativa. Para nuestro modelo, se tienen en cuenta únicamente las contribuciones entre residuos i y j que estén en fragmentos peptídicos distintos. Las interacciones de un residuo con los demás del mismo fragmento no se calculan debido a que, al tratarse de cuerpos rígidos, su contribución es igual en todas las conformaciones. El parámetro de peso ω_{ij} vale 1 si en la conformación nativa se define un contacto entre los residuos i y

j, y vale 0 si no existe contacto.

En un estudio preliminar, definimos por exceso el número de contactos presentes en la conformación nativa[99]. Entonces asumimos que los aminoácidos i y j están interaccionando en la conformación nativa cuando la distancia d_{ij}^{PDB} entre sus carbonos-α es menor o igual a una distancia umbral, que establecimos en un valor muy alto. De esta manera se obtiene un número muy elevado de contactos para cada proteína, y se genera una superficie de búsqueda de gran rugososidad, con mínimos locales separados por barreras energéticas muy grandes.

En esta memoria presentamos una definición más refinada de los contactos nativos para el modelo de Gō[119]. Para un par ij de residuos, en lugar de definir los contactos a partir de la distancia entre sus carbonos-α en la conformación nativa, utilizamos la menor de las distancias entre cualesquiera de sus átomos pesados $d_{a_i b_j}^{PDB}$, donde a_i y b_j son los átomos más próximos del par de residuos. Definimos un contacto cuando la distancia $d_{a_i b_j}^{PDB}$ es inferior a 4.5 Å. Este valor de referencia es algo superior al doble del promedio de radios de Van der Waals para grupos atómicos de proteínas[120], y ha sido utilizado con éxito en otros estudios[121].

Como hemos comentado, utilizamos el valor de la energía global como función de mérito del algoritmo. Para una proteína, esta función tiene su mínimo en 0, valor correspondiente a su conformación nativa. La energía de cualquier otra conformación siempre tiene valor positivo y finito, por ser suma de términos E_{ij} positivos y finitos. Estos términos pueden alcanzar valores muy altos en el caso de que las topologías que exploremos sean muy diferentes de la nativa. Cuando el número de fragmentos es medio o alto, se genera un complejo espacio de búsqueda que favorece conformaciones compactas, a menudo difíciles de explorar por métodos como el de Monte Carlo. Por tanto, esta definición de potencial puede suponer una buena prueba para el algoritmo genético.

3.2. Minimización de la energía para proteínas todo α

Para poner a prueba nuestra estrategia evolutiva, hemos llevado a cabo la búsqueda del empaquetamiento de fragmentos de mínima energía para un conjunto de proteínas con el potencial de tipo Gō. En la Figura 3.1 mostramos el conjunto de proteínas con el que realizamos este estudio. Por analogía con el estudio de potenciales hidrófobos —que ocupa el siguiente capítulo—, hemos seleccionado proteínas todo α de distinto grado de complejidad del Protein Data Bank. Las proteínas seleccionadas tienen entre 3 y 9 hélices α, con lazos de distinto tamaño. Vamos a considerar las hélices α de estas proteínas como fragmentos rígidos de nuestro estudio. El distinto número de fragmentos repercute en la complejidad del muestreo porque el número de variables que el algoritmo tiene que optimizar depende del número de fragmentos.

Para cada una de las proteínas, tomamos la secuencia y las coordenadas de sus átomos del PDB. En esta parte de nuestro estudio únicamente consideramos aquellos residuos que se encuentran en hélices α de la estructura nativa de la proteína, según las indicaciones sobre estructura secundaria del encabezamiento del archivo PDB. Como ya hemos comentado, las hélices individuales son los fragmentos rígidos. Por tanto el muestreo conformacional se realiza sobre los posibles empaquetamientos de los fragmentos. Los residuos que se encuentran en los lazos entre hélices los contabilizamos únicamente para determinar la máxima separación entre dos fragmentos sucesivos en la secuencia de la proteína. El número de contactos nativos se define como hemos descrito en la Sección 3.1 para los residuos que forman parte de los fragmentos. En la Figura 3.1 mostramos el número de contactos nativos para cada una de las proteínas, que en general es creciente con el número de fragmentos de la proteína.

Para todas las proteínas hemos llevado a cabo los cálculos correspondientes a las tres codificaciones que hemos descrito en la Sección 2.2: la externa, la interna simple

Descripción	Código PDB	nº fragmentos rígidos	nº residuos en el modelo	nº enlaces peptídicos en lazos	nº contactos nativos	
Proteína transportadora de apo-D-alanina	**1dv5**	3	40	20 - 19	17	
Anafilotoxina c5a porcina	**1c5a**	4	49	5 - 7 - 4	41	
Dominio SAM del receptor EPH4	**1b0x**	5	51	4 - 5 - 3 - 5	42	
Subunidad δ de la ATP sintasa F1F0	**1abv**	6	79	4 - 3 - 6 - 6 - 7	74	
Dominio Death de IRAK-1	**1wh4**	7	75	2 - 3 - 10 - 4 - 2 - 2	96	
Dominio regulador de señalización de proteína G	**1ezt**	8	105	2 - 2 - 2 - 2 - 11 - 8 - 2	96	
Dominio VI de la calpaína	**1nx2**	9	122	2 - 3 - 11 - 8 - 11 - 5 - 7 - 10	97	

Figura 3.1: Conjunto de proteínas para el estudio con el potencial de tipo Gō, con el número de fragmentos en el modelo, el número de residuos, el número de enlaces peptídicos en los lazos, el número de contactos nativos en el potencial y una representación de su estructura tridimensional.

Parámetros de la minimización	
tamaño de la población	100
nº generaciones por ciclo	500
p_{cross}	0.5
p_{mut}	0.1
$RMSD_{ij}^{min}$	1. Å
nº optimizaciones independientes	10

Tabla 3.1: Parámetros de la estrategia evolutiva para la minimización de la energía con el potencial de tipo Gō.

y la interna compleja. Como comentamos entonces, la codificación interna compleja y la interna simple son complementarias, por lo que las hemos utilizado conjuntamente. Cada uno de los cromosomas de la población se decodifica de acuerdo con los dos métodos, lo que origina dos conformaciones diferentes. Se calcula la energía para estas dos conformaciones, y consideramos que la conformación que corresponde al cromosoma es la de menor energía de las dos. Por tanto, en nuestros resultados hablamos únicamente de las codificaciones externa e interna.

En la Tabla 3.1 mostramos los parámetros que hemos utilizado para llevar a cabo las minimizaciones energéticas con la estrategia evolutiva. Cada uno de los experimentos lo realizamos 5 veces, modificando únicamente el conjunto de números semilla para el generador de números aleatorios.

3.3. Resultados y discusión

En la Tabla 3.2 mostramos los resultados de las minimizaciones con el potencial de Gō para cada una de las codificaciones descritas en la Sección 2.2. Lo primero que queremos destacar en los resultados es que la eficiencia del algoritmo depende de la complejidad del problema de búsqueda. Como era de esperar, para problemas de búsqueda sencillos o de mediana dificultad, el método es muy eficaz. Con ambas codificaciones, el valor de

Codificación:		**Interna**		**Externa**	
PDBid	*nº frag*	$E_{min}/nº\ cont$	*RMSD* (Å)	$E_{min}/nº\ cont$	*RMSD* (Å)
1dv5	3	$9.46 \cdot 10^{-3}$	0.64	$4.31 \cdot 10^{-5}$	0.02
1c5a	4	$5.06 \cdot 10^{-3}$	0.13	$7.60 \cdot 10^{-5}$	0.03
1b0x	5	$8.09 \cdot 10^{-3}$	0.14	$3.16 \cdot 10^{-5}$	0.10
1abv	6	$8.37 \cdot 10^{-2}$	0.91	$2.57 \cdot 10^{-3}$	0.11
1wh4	7	$2.50 \cdot 10^{-1}$	2.57	$9.01 \cdot 10^{-3}$	0.17
1ezt	8	$9.29 \cdot 10^{-1}$	10.16	$8.60 \cdot 10^{-1}$	5.11
1nx2	9	$6.60 \cdot 10^{-1}$	6.87	$9.11 \cdot 10^{-1}$	8.87

Tabla 3.2: Resultados de energía mínima normalizada y *RMSD* respecto a la conformación nativa en angstroms para las dos codificaciones, obtenidos con la estrategia evolutiva para el conjunto de proteínas.

RMSD del mínimo frente a la conformación nativa es muy próximo a 0 para proteínas de hasta 6 fragmentos. Con proteínas de mayor tamaño, los resultados empeoran notablemente. Como puede observarse en los valores de la Tabla 3.2, la energía mínima aumenta con las dos codificaciones al aumentar el tamaño de la proteína. Esta pérdida de eficiencia se debe a que cuando tratamos un número elevado de fragmentos, el algoritmo debe ajustar simultáneamente muchas variables. En el caso de mayor complejidad, la proteína cuyo código del PDB es 1nx2, se intentan optimizar hasta 48 variables simultáneamente [$6 \times (9-1)$ fragmentos]. Es normal, por tanto, que la bondad de los resultados esté relacionada con el tamaño de la proteína considerada.

A pesar de la pérdida de eficiencia del algoritmo, dependiente del tamaño de la proteína, también para las proteínas de mayor complejidad los resultados son razonablemente buenos. Veamos, por ejemplo, el caso de 1ezt, una proteína reguladora de la señalización de proteínas G, con 8 hélices α en la estructura tomada del PDB. En la Figura 3.2 mostramos los mapas de distancias para esta proteína en sus conformaciones nativa y optimizada con el potencial de Gō. En este caso, el resultado procede de la optimización con la codificación interna. El valor normalizado de la energía para la conformación optimizada es de 0.929, y su *RMSD* respecto a la conformación nativa

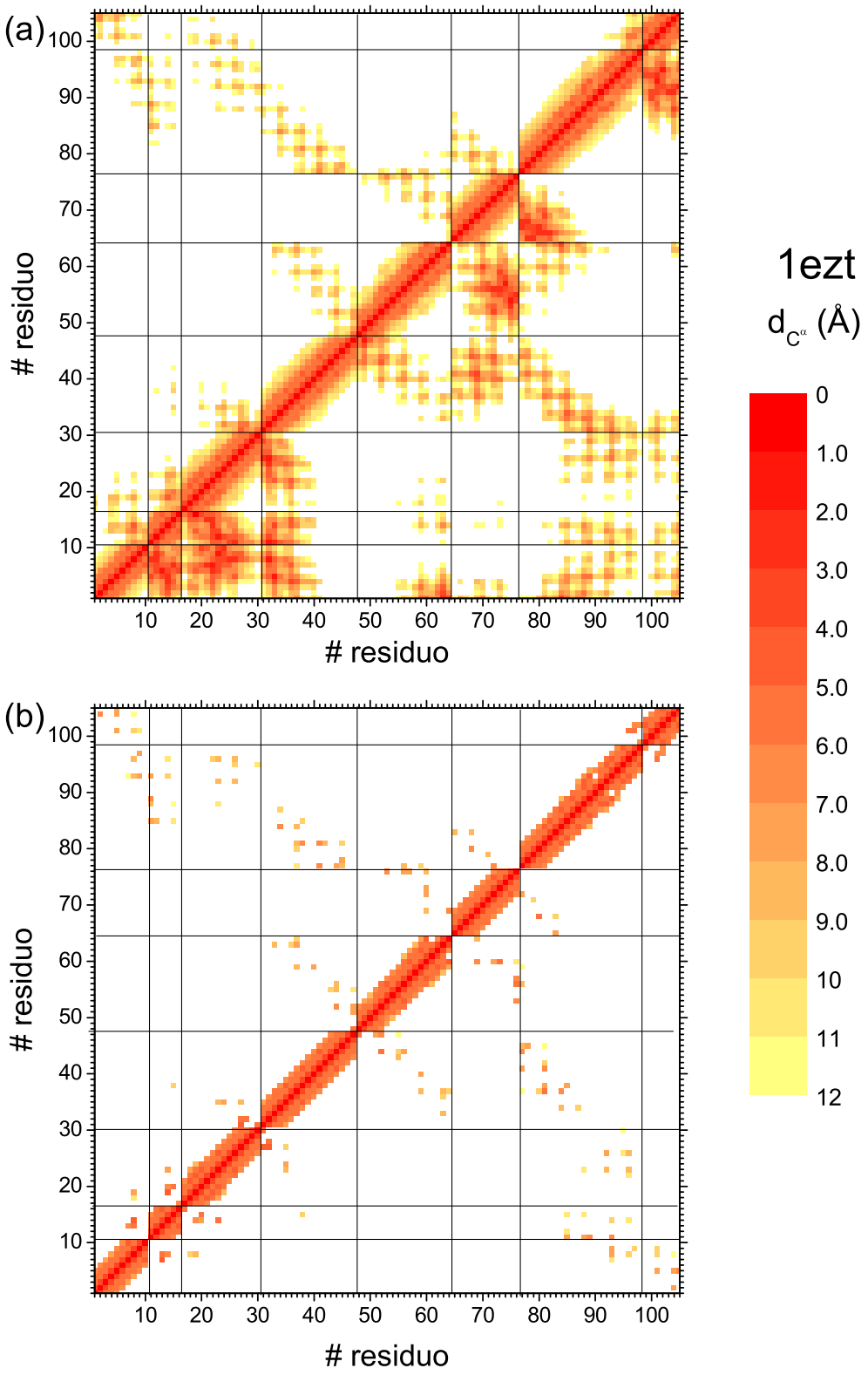


Figura 3.2: Mapas de distancias de 1ezt. (a) Mapa para todos los residuos de la proteína considerados en el modelo reducido. (b) Mapa para residuos con ω_{ij}=1 en el potencial de Gō. En ambos, representamos distancias en la conformación nativa —triángulo superior— y en la optimizada con la estrategia evolutiva con codificación interna —triángulo inferior—. Distancias entre carbonos α en angstroms.

asciende a 10.16 Å. Se trata del peor de los resultados que hemos obtenido para las dos codificaciones. En el primero de los mapas, que mostramos en la Figura 3.2 (a), se representan las distancias entre carbonos-α para todos los residuos de las ocho hélices de la proteína. En el triángulo superior representamos distancias en la conformación nativa y en el inferior, en la conformación optimizada. Si comparamos ambos triángulos, encontramos regiones donde hay grandes diferencias entre ambas conformaciones. La optimizada es mucho más compacta que la nativa, como podemos deducir por la trama de contactos, mucho más tupida en el triángulo inferior. En la Figura 3.2 (b) mostramos el mapa de distancias para las mismas conformaciones de 1ezt, pero se representan únicamente las distancias entre pares que cuentan a la hora de calcular la energía, es decir, aquellos pares ij para los cuales ω_{ij}=1. En este mapa, al contrario de lo que sucede con el primero, apenas podemos observar diferencias entre la conformación nativa y la optimizada, con una precisión de 1 Å. Esto significa que si consideramos únicamente los residuos que definen la conformación nativa, la estructura está bastante bien ajustada. El algoritmo es razonablemente eficaz realizando su búsqueda en una superficie de energía compleja, aunque no sea capaz de refinar suficientemente los buenos resultados obtenidos.

Hemos comprobado el efecto que tiene el número de contactos nativos que se definen en el potencial de Gō. Para la proteína 1ezt, hemos modificado la distancia umbral para la definición de un contacto ($d_{a_ib_j}$, ver Sección 3.1), aumentándolo de 4.5 a 7 Å. Esta nueva distancia de corte permite que se definan hasta 310 contactos en la conformación nativa, en lugar de los 96 que se definían con 4.5 Å. Mostramos los resultados de la optimización para ambos valores de $d_{a_ib_j}$ en la Tabla 3.3. Para un valor de $d_{a_ib_j}$=7 Å obtenemos valores de energía mínima normalizada y de $RMSD$ notablemente más bajos que para una distancia de corte de 4.5 Å, con ambas codificaciones. Esto se debe a que, al aumentar el número de contactos nativos, la superficie de energía está

Codificación:		Interna		Externa	
$d_{a_ib_j}$ (Å)	$n^o cont$	$E_{min}/n^o cont$	$RMSD$ (Å)	$E_{min}/n^o cont$	$RMSD$ (Å)
4.5	96	$9.29 \cdot 10^{-1}$	10.16	$8.60 \cdot 10^{-1}$	5.11
7.0	310	$5.82 \cdot 10^{-2}$	0.55	$2.68 \cdot 10^{-2}$	0.40

Tabla 3.3: Resultados de energía mínima normalizada y $RMSD$ respecto a la conformación nativa en angstroms, obtenidos con la estrategia evolutiva para la proteína 1ezt, con las dos codificaciones, con distinto valor de la distancia de corte $d_{a_ib_j}$ para los contactos nativos en el potencial de Gō.

mejor definida: aparecen pozos más profundos y barreras entre mínimos locales más altas. En esta superficie más rugosa pero con un mínimo absoluto más pronunciado, la optimización profundiza mejor en los mínimos energéticos una vez se han localizado. Esto permite alcanzar valores más bajos de energía normalizada que corresponden a conformaciones mucho más parecidas a la nativa.

Otra conclusión que podemos extraer de los valores de energía mínima de la Tabla 3.2 es que, para un mismo número de contactos nativos, el algoritmo funciona mejor cuanto menor es el número de fragmentos. Las proteínas 1wh4, 1ezt y 1nx2 tienen aproximadamente el mismo número de contactos. La energía normalizada es más baja para 1wh4 con las dos codificaciones que para 1ezt y 1nx2. Los mismo sucede con los valores de $RMSD$, también para las dos codificaciones. Esto también resulta fácil de comprender. El número de contactos es el número de restricciones que definen la superficie de energía. Para un mismo número de restricciones, cuanto mayor es el número de variables a optimizar más difícil resulta encontrar la conformación de mínima energía. Este tipo de comportamiento lo encontramos más acusadamente en el caso de la codificación externa.

A menudo sucede que varios experimentos de minimización para una proteína con el mismo conjunto de parámetros, exceptuando los números semilla para la generación de números aleatorios, no alcanzan el mismo resultado final. Esta falta de convergencia en los valores finales se hace más probable cuanto más complicada es la superficie

Codificación:		**Interna**	**Externa**
$PDBid$	$n^o frag$	S_E	S_E
1dv5	3	0.02	0.0002
1c5a	4	0.03	0.0002
1b0x	5	0.2	0.002
1abv	6	0.4	0.002
1wh4	7	0.3	0.5
1ezt	8	0.5	0.5
1nx2	9	1.	0.2

Tabla 3.4: Desviaciones estándar de los valores de energía mínima obtenidos para los distintos experimentos de optimización llevados a cabo con la estrategia evolutiva para las dos codificaciones con el potencial de Gō.

de muestreo. Esto es lo que observamos en los resultados de nuestros experimentos si atendemos a la convergencia de las distintas optimizaciones. En la Tabla 3.4 mostramos los valores de la desviación estándar de los valores finales de la energía de las distintas optimizaciones realizadas para cada proteína con un mismo conjunto de parámetros con las dos codificaciones. En general, los resultados que hemos obtenido para las proteínas con menos fragmentos son más reproducibles que los que ofrece el algoritmo para proteínas con más fragmentos. Para ambas codificaciones, la desviación estándar de la energía mínima de las distintas minimizaciones para cada proteína aumenta al incrementarse el número de fragmentos que tratamos. Al crecer el tamaño de la proteína, el algoritmo tiene que optimizar más variables en una superficie de dimensionalidad cada vez mayor. Esto provoca un estancamiento en la minimización, a partir del cual le resulta muy difícil refinar el valor alcanzado para la energía.

Ponemos como ejemplo de este comportamiento el mejor y el peor de los resultados que hemos obtenido con la codificación interna para una misma proteína, 1b0x, un dominio de un receptor tirosina quinasa. Se trata de una proteína con 5 hélices α en su conformación nativa, por lo que en nuestro modelo la dividimos en cinco fragmentos. En la mejor de las ejecuciones independientes de la estrategia evolutiva, la conformación

final tiene $RMSD$=0.14 Å y E=8.09·10^{-3}. En el peor de los casos, la conformación final tiene $RMSD$=5.33 Å y E=5.88·10^{-1}. En la Figura 3.3 mostramos los mapas de distancias entre carbonos-α para ambas conformaciones. Como en los que hemos mostrado antes para 1ezt, cada uno de los mapas contiene información correspondiente a la conformación nativa en el triángulo superior, y a la conformación optimizada en el inferior. En los dos mapas de la parte superior de la Figura, (a) y (b), representamos las distancias entre todos los carbonos-α en el modelo de la proteína. En cambio, en los de la parte inferior de la Figura, (b) y (d), sólo se muestran distancias entre residuos interaccionantes. Estos últimos son prácticamente iguales para las dos conformaciones minimizadas, la de $RMSD$=5.33 Å (c), y la de $RMSD$=0.14 Å (d). En ambos casos apenas se observa diferencia con la conformación nativa, con una precisión de 1 Å. Sin embargo, si atendemos a los mapas (a) y (b) de la Figura 3.3, en los que mostramos distancias entre todos los residuos de los fragmentos, lo que observamos es muy diferente. Para el mejor de los resultados (a), la conformación minimizada es prácticamente igual a la nativa. Para el peor (b), las diferencias entre nativa y optimizada son muy importantes. La conformación con $RMSD$=5.33 Å tiene un empaquetamiento de hélices mucho más denso. Pensamos que esto se debe a que la optimización de las variables de los cromosomas se produce en dos "etapas". En primer lugar, se ajustan en cierta medida las distancias correspondientes a regiones con muchos pares cuyo ω_{ij}=1, es decir, aquellos que están interaccionando más fuertemente con el potencial de Gō. Esta primera "etapa" de la optimización la cumplen todos los experimentos para cada una de las proteínas. A continuación, el algoritmo empaqueta las hélices en las regiones en que hay menos interacciones definidas. Por tanto, entre una conformación final buena y una muy buena se requiere una labor de refinado en la optimización que no siempre se lleva a cabo. Esa labor de refinado correspondería al correcto ajuste de las regiones donde hay menos contactos definidos. Los resultados mostrados anteriormente para la proteína 1ezt con

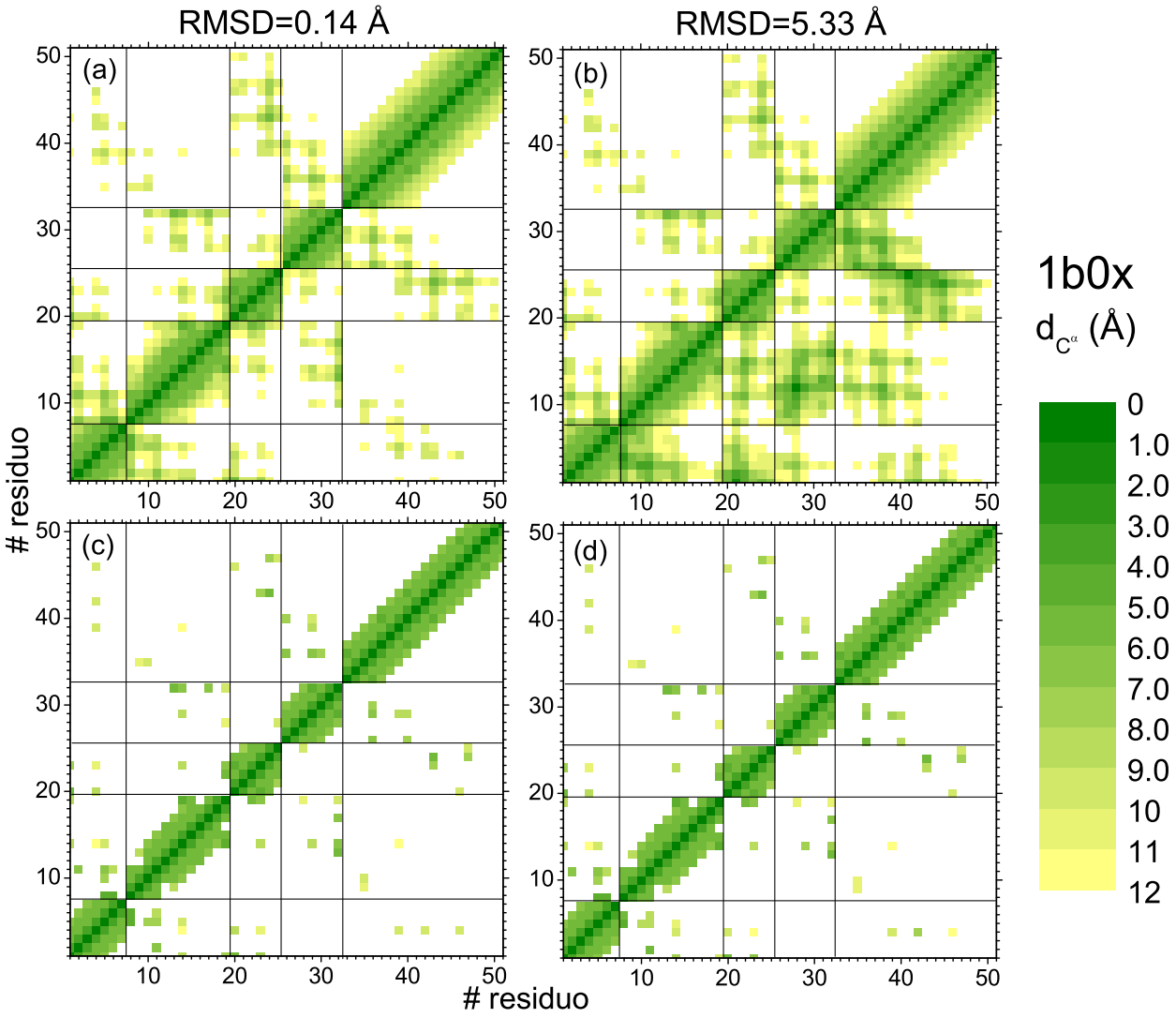


Figura 3.3: Mapas de distancias de dos conformaciones finales obtenidas para 1b0x con la estrategia evolutiva con la codificación interna: (a) y (c), $RMSD$=0.14 Å; (b) y (d), $RMSD$=5.33 Å. (a) y (b): mapas con todos los residuos considerados en el modelo reducido; (c) y (d): mapas para residuos con ω_{ij}=1 en el potencial de Gō. En todos los mapas, conformación nativa en el triángulo superior, y optimizada en el inferior. Distancias entre carbonos α en Å.

una distancia de corte más alta en la definición de contactos (Tabla 3.3) indican que el algoritmo tiende a realizar este refinamiento cuando el potencial es más restrictivo.

Como hemos ido viendo a lo largo de esta sección, la estrategia evolutiva no funciona igual con las dos codificaciones. En primer lugar, porque con la codificación externa

se obtienen casi siempre mejores resultados que con la interna (ver Tabla 3.2). Para las proteínas de menor tamaño hay hasta dos órdenes de magnitud entre los valores de mejor energía alcanzados con la codificación externa y la interna. En segundo lugar, con la codificación externa se alcanzan resultados más reproducibles que con la interna en casi todos los casos (ver Tabla 3.4). Para una misma proteína, con esta codificación la labor de refinado se realiza mucho más eficazmente. Esto tiene que ver con que las optimizaciones son, en casi todos los casos, sensiblemente más largas con la codificación externa. En la Figura 3.4 mostramos una representación del número de generaciones que tarda la minimización en alcanzar el mínimo correspondiente para las dos codificaciones. Observamos que, para proteínas con entre 3 y 9 fragmentos, las optimizaciones realizadas con la codificación interna tardan aproximadamente lo mismo. En cambio, con la codificación externa la duración de los cálculos crece linealmente con el tamaño para proteínas con hasta 6 fragmentos. Podemos correlacionar esto con que, al aumentar la complejidad del problema de búsqueda, para la codificación externa se mantiene la eficiencia del método. Sin embargo, los resultados empeoran cuando utilizamos la interna. Para los casos con un mayor número de fragmentos, con la codificación interna se tarda aproximadamente lo mismo, mientras que con la externa las fluctuaciones son mayores.

Como hemos comentado, en la codificación interna cada cromosoma tiene dos significados, el procedente de la interna simple y el de la interna compleja. Hemos observado que en las optimizaciones en las que se manejan pocos fragmentos, la población se reparte entre individuos de ambas codificaciones. Sin embargo, para proteínas de mayor tamaño, en general la población se agrupa fundamentalmente en un solo tipo de codificación[99]. Esto puede deberse a que se hace cada vez más difícil encontrar un buen empaquetamiento de todos los fragmentos. Una vez se ha localizado, la población concentra su actividad en refinar los dominios bien formados de esa estructura.

No podemos descartar que un ajuste más refinado de los parámetros mejore la

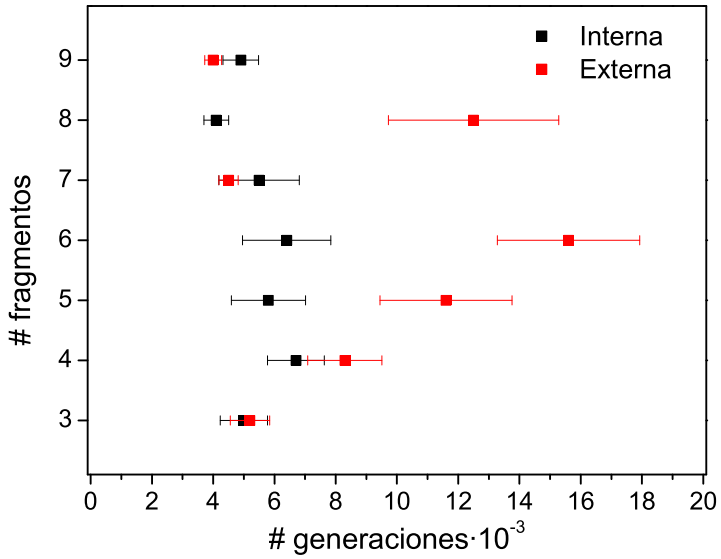


Figura 3.4: Duración en número de generaciones de la optimización realizada con la estrategia evolutiva para las dos codificaciones en función del número de fragmentos.

eficiencia de la minimización con la codificación interna. Hemos comprobado que sus mejores resultados se obtienen cuando se mantienen subpoblaciones de las dos variantes de la codificación, simple y compleja[99]. Esto podría permitir algún grado de transferencia de conformaciones entre ambas. Por otra parte, la diversidad de la población está relacionada con el umbral de $RMSD$ entre conformaciones de la población. Pudiera ser que hubiese que ajustar esa diferencia mínima al tamaño de la proteína, ya que esa magnitud es sensible al número de centros que se contabilizan para calcularla. Pero esto podría mejorar los resultados también para la codificación externa. De modo que, en términos generales, podemos considerar que ésta está funcionando mejor que aquella. Sin embargo, cuál se utilice finalmente dependerá de las particularidades del tipo de potencial que estemos analizando.

3.4. Resumen del Capítulo y conclusiones

En este Capítulo hemos explicado la puesta a punto de nuestro método evolutivo para el estudio de potenciales de plegamiento de proteínas. En este caso, en lugar de utilizar un potencial de interacción realista, empleamos una función de potencial de tipo Gō[98] como función de mérito del algoritmo. Se trata de un potencial *ad hoc*, específico para cada proteína. Cualquiera que sea la proteína con la que trabajemos, este tipo de potencial tiene su mínimo energético definido, por construcción, en su conformación nativa. Si consideramos otra conformación de la misma proteína, su energía se calcula en función de sus diferencias con la nativa. En nuestro caso, para calcular la energía basándonos en estas diferencias, hemos utilizado una definición de los contactos atómicos que aparecen en la conformación nativa[121]. Para aquellos pares de residuos entre los que aparece algún contacto atómico, definimos un término en el potencial de Gō, dependiente de la distancia entre los carbonos-α de los dos residuos. Finalmente, la energía de una conformación se calcula sumando las contribuciones de cada uno de los pares interaccionantes. Tal y como lo hemos definido, este potencial engendra un complejo campo de búsqueda, semejante al de los potenciales de interacción realistas. Por ello supone un test óptimo para nuestra metodología, a pesar de que el potencial en sí carezca de ninguna capacidad predictiva.

Para la puesta a punto del método con el potencial de Gō hemos seleccionado un conjunto de proteínas de tipo todo α. Como explicamos en el Capítulo 2, cuando utilizamos proteínas de este tipo, que también podemos utilizar para evaluar potenciales hidrófobos, los fragmentos rígidos de nuestro método son hélices α. En este caso, no consideramos las interacciones de los residuos que se encuentran en los lazos entre hélices. Hemos seleccionado proteínas con distinto número de hélices α en su conformación nativa. Esto permite comprobar la variación en la eficiencia del método con la

complejidad del problema de búsqueda. Además, hemos estudiado la dependencia de los resultados con las codificaciones que permiten transformar conformaciones de la proteína en cromosomas del algoritmo.

Como hemos visto, el algoritmo resulta muy eficiente en la búsqueda del mínimo energético para el conjunto de proteínas que hemos estudiado. Para todos los casos se alcanzan valores muy próximos al mínimo de la función de mérito. Aun así, hemos comprobado que la eficiencia del método depende de una serie de factores. En primer lugar, el método encuentra el mínimo con más dificultad cuanto mayor es el número de fragmentos rígidos que tiene que empaquetar. Esto es razonable porque el número de fragmentos está directamente relacionado con el número de variables en cada uno de los cromosomas que el algoritmo tiene que ajustar para alcanzar el mínimo. Sin embargo, hemos comprobado que la eficiencia del método, para aquellas proteínas que ofrecen resultados más deficientes, depende muy fuertemente de la buena definición de la superficie de energía. Para una de estas proteínas, hemos observado que al aumentar el número de contactos nativos mejoran mucho los resultados. Esta mejora se debe a que los contactos nativos contribuyen a definir mejor la superficie de energía, y que cuantos más haya, mejor se puede alcanzar el mínimo energético con nuestro método. Por tanto, hemos comprobado que para una superficie de energía muy bien definida, la estrategia evolutiva es capaz de encontrar una conformación muy próxima a la nativa.

En su libro sobre algoritmos genéticos[64], Goldberg retó a la comunidad que trabaja con algoritmos evolutivos a generar distintas codificaciones para transformar las soluciones de su problema de búsqueda en cadenas de variables. Para nuestro trabajo hemos elaborado hasta tres codificaciones del problema, dos de las cuales trabajan coordinadamente por ser complementarias (ver Capítulo 2). En este Capítulo hemos evaluado el comportamiento de las codificaciones externa e interna. Los mejores resultados los hemos obtenido con la codificación externa. Esto puede estar determinado por las

pequeñas diferencias en la definición de la función de mérito. Como hemos comentado, con la codificación externa introducimos penalizaciones para algunas conformaciones en función de la distancia entre residuos contiguos del modelo de la proteína. Dado que estas penalizaciones se suman a la energía, la función de mérito no es exactamente igual en ambas codificaciones. Esto puede afectar en los primeros estadios de la optimización, cuando el muestreo con la codificación externa genere muchas conformaciones susceptibles de ser penalizadas. Sin embargo, tras las primeras etapas de la optimización, las dos codificaciones estarían en igualdad de condiciones.

Nos inclinamos a pensar que es por otro motivo que aquella funciona mejor. Goldberg definía una serie de principios para elaborar una codificación[64]. Uno de estos principios hace referencia a la relación entre diferentes segmentos de un cromosoma, que pueden transferirse eventualmente mediante las operaciones de entrecruzamiento. En concreto, Goldberg sugería que, dentro de la cadena de variables, cada segmento debía ser lo más independiente posible de los demás. La codificación interna, al construir las conformaciones, utiliza de manera combinada los grupos de variables —cada fragmento utiliza como sistema de referencia el anterior—. En cambio, la externa utiliza cada grupo de variables independientemente. De modo que podría ser por esta interrelación entre variables que la optimización es menos eficiente con la codificación interna. Esto podría explicar por qué, con la codificación interna, una vez encontrado un mínimo, se refina peor que con la externa. Pensemos, por ejemplo, en una mutación en la que se modifica una variable en un cromosoma que corresponde a una conformación bien empaquetada. En la codificación externa el cambio en uno solo de los fragmentos debido a la mutación puede mejorar la energía. Sin embargo, en la codificación interna, muchos fragmentos pueden verse afectados por el cambio en una variable. Esto puede hacer más difícil el refinado de los mínimos en la codificación interna que en la externa.

El estudio incluido en este Capítulo amplía el que habíamos publicado previamen-

te[99]. En conjunto, hemos podido conocer en profundidad las capacidades de nuestro método evolutivo para la minimización. Para las distintos potenciales y tipos de proteínas con que trabajemos, habrá que realizar pequeñas modificaciones en la metodología. Además, en función del sistema que estudiemos, puede ser más conveniente utilizar una u otra codificación. Sin embargo, en términos generales, la estrategia evolutiva tal y como la hemos descrito queda preparada para la evaluación de modelos de interacción más realistas.

Capítulo 4

Evaluación de potenciales hidrófobos

En el Capítulo anterior hemos descrito la evaluación de nuestro método evolutivo para la minimización de la energía de proteínas con una representación de fragmentos rígidos. Hemos comprobado que para problemas de búsqueda más o menos sencillos el algoritmo evolutivo funciona muy eficientemente. Con nuestro método podemos encontrar el mínimo energético para una proteína cuando éste es profundo y la superficie de energía está bien definida. En este Capítulo de la Tesis Doctoral utilizamos nuestro algoritmo evolutivo para estudiar una serie de potenciales de los denominados "basados en estructuras", definidos para intentar reproducir la interacción hidrófoba.

Hay varias motivaciones por las cuales muchos grupos de investigación han dedicado sus esfuerzos a obtener potenciales basados en estructuras[80–96]. Como hemos comentado en el Capítulo 1, una descripción atómica de la proteína con una definición detallada de sus interacciones es más rigurosa que una descripción de grano grueso. Sin embargo, los potenciales de interacción atómicos tienen aplicabilidad limitada para procesos como el plegamiento de proteínas, debido a su escala de tiempo. Por este motivo, los modelos de menor resolución con potenciales de interacción sencillos son más apropiados para la simulación del plegamiento. Por otra parte, hay un gran número de

estructuras de proteínas resueltas por cristalografía de rayos-X y resonancia magnética nuclear (RMN), disponibles en el Protein Data Bank[56,57]. Esto proporciona una gran cantidad de información acerca de las propensiones de los distintos aminoácidos a estar en un determinado entorno de la proteína. Los potenciales basados en estructuras se obtienen transformando estas propensiones en valores de energía de interacción[54].

En la mayoría de casos, los potenciales de interacción que se han desarrollado son potenciales estadísticos de campo medio. Estos potenciales se deducen a partir de la información que proporciona un conjunto de estructuras utilizando la ley de Boltzmann. Si queremos calcular el potencial de campo medio $u_{ij}(r)$ entre dos centros de interacción i y j situados a una distancia comprendida en el intervalo $(r-\Delta r)$ y $(r+\Delta r)$ podemos usar una ecuación del tipo[54]

$$u_{ij}(r) = -RT \ln \frac{N_{ij}^{obs}(r)}{N^{ref}(r)} \ , \tag{4.1}$$

donde R es la constante de los gases y T es la temperatura. En esta ecuación, N_{ij}^{obs} es el número de veces que los centros i y j aparecen a una distancia que esté en el intervalo $(r-\Delta r, r+\Delta r)$, contabilizado en el conjunto de proteínas seleccionado. $N^{ref}(r)$ es el número de pares en el mismo intervalo en el estado de referencia. La diferencia entre los distintos potenciales estadísticos estriba en factores como la definición de este estado de referencia, los centros de interacción seleccionados o la resolución en la dependencia con la distancia entre centros.

Uno de los estados de referencia más utilizados es el de la llamada aproximación cuasiquímica de Bethe[122]. Según esta aproximación, las interacciones entre residuos de la proteína son del mismo tipo que las que se dan en una mezcla de aminoácidos libres y moléculas de disolvente[88]. En cuanto a los centros de interacción, la aproximación más común es seleccionar un sólo centro por residuo de aminoácido[54], aunque también pueden

utilizarse descripciones más detalladas. Por último, con respecto a la dependencia con la distancia, los potenciales más sencillos son los denominados potenciales de contacto. Estos potenciales se calculan a partir del número de veces que dos centros de interacción aparecen a una distancia menor que un umbral. La forma de estos potenciales es de pozo cuadrado: el valor de su energía para un par de centros ij es cero a distancias superiores a ese umbral y u_{ij} a distancias inferiores. Estos potenciales resultan muy útiles, por ejemplo, para realizar simulaciones en red. Otra opción es obtener potenciales con una dependencia con la distancia más detallada, más apropiados para simulaciones fuera de red. En estos casos, el intervalo de distancias de interés se divide en una serie de segmentos. En función del número de contactos contabilizados en los distintos segmentos, el potencial tendrá un valor diferente.

Una alternativa para obtener potenciales a partir de estructuras, menos habitual que la aproximación de campo medio, es utilizar métodos matemáticos[84,92,93]. En este tipo de método, se requiere en primer lugar un gran número de quimeras o conformaciones alternativas de la proteína. Estas quimeras se generan mediante el método de "threading" o enhebrado, que consiste en introducir la secuencia de una proteína en la estructura de otras proteínas. Así, una secuencia S_n, cuya conformación nativa viene dada por las coordenadas X_n, se introduce en las estructuras X_j para un conjunto de J proteínas. Para todas las combinaciones de N secuencias y J estructuras, se calcula la energía en función de un conjunto de incógnitas P. Estas incógnitas son precisamente los términos energéticos entre pares para una serie de intervalos de distancias. Siguiendo la hipótesis termodinámica de Anfinsen[4], para una secuencia dada la energía de la conformación nativa $E(S_n, X_n; P)$ corresponde al mínimo energético. Por tanto, la energía de cualquiera de las quimeras $E(S_n, X_{j\neq n}; P)$ tiene que ser superior a la de la nativa. Así, con las funciones para la energía se plantean un gran número de inecuaciones en

las que se impone la condición

$$E(S_n, X_{j\neq n}; P) - E(S_n, X_n; P) > 0,\ j = 1, ..., J;\ n = 1, ..., N, \tag{4.2}$$

La solución de las inecuaciones conduce al potencial de interacción que hace ciertas las proposiciones que responden a la Ecuación 4.2.

Normalmente, al desarrollo de los potenciales le sigue una evaluación, para comprobar su capacidad de definir un mínimo global profundo para la energía y si este mínimo corresponde a la conformación nativa. En general, para esta evaluación se utilizan métodos basados en quimeras[97]. Concretamente, la aproximación es la inversa a la utilizada para generar potenciales resolviendo inecuaciones. Para un conjunto de secuencias, se genera un gran número de quimeras por enhebrado de secuencias en estructuras, dinámica molecular o enumeración exhaustiva. A continuación, se calcula la energía de cada una de las quimeras de acuerdo con el potencial. La energía de todas estas quimeras tiene que ser superior a la de la conformación nativa para cada secuencia. A menudo se realizan medidas del denominado "z-score", que permite estimar cuánto difiere el valor de la nativa con respecto a las energías para las quimeras. Este tipo de evaluaciones son útiles porque permiten comparar un gran número de estructuras plegadas, pero no permiten evaluar un potencial como con simulaciones del plegamiento o en experimentos de minimización. En estos otros tipos de estudio se genera un gran número de conformaciones de la proteína siguiendo el camino que dicta el propio potencial. De este modo, el muestreo que se puede realizar sobre el espacio conformacional de la proteína es más amplio que utilizando exclusivamente quimeras. Por este motivo, aplicamos nuestro método de minimización para evaluar potenciales de plegamiento basados en estructuras.

En nuestro estudio, hemos querido evaluar potenciales de interacción tomados de la bibliografía, representativos de las distintas aproximaciones. Así, hemos seleccionado los

potenciales que denominamos de Nanias[100], TE-13[92,93] y DFIRE-SCM[96,123]. El potencial de Nanias es el más sencillo de los tres. Está basado en un potencial de contacto, la matriz de energías que Miyazawa y Jernigan[88] obtuvieron utilizando la aproximación cuasiquímica. A este potencial, Nanias *et al.* le añaden una función sencilla para la dependencia con la distancia. El segundo de los potenciales que hemos tomado de la bibliografía es el TE-13 de Tobi y Elber[92,93]. No es un potencial estadístico, sino que se construye mediante la resolución de un gran número de inecuaciones. Este potencial tiene una dependencia con la distancia mucho más detallada que el de Miyazawa y Jernigan. El tercer potencial que consideramos en el estudio es el DFIRE-SCM de Zhou *et al.*[96,123]. Este potencial es de tipo estadístico como la matriz de energías de contacto de Miyazawa y Jernigan. Zhou *et al.* utilizan un nuevo estado de referencia de "gas ideal finito" que parece mejorar los resultados de otros potenciales[123]. Como el TE-13, tiene una dependencia con la distancia entre centros dividida en una serie de intervalos. Más adelante describiremos en mayor profundidad cada uno de estos tres potenciales de interacción.

Los potenciales basdados en estructuras representan la interacción efectiva entre pares de residuos en un "baño" de aminoácidos densamente empaquetados[54]. Por tanto, una componente del potencial representa la tendencia a empaquetarse de los residuos y otra representa las tendencias específicas entre pares de aminoácidos según el tipo al que correspondan en el seno de la proteína. Para separar estos dos efectos, de colapso y de secuencia, hemos querido hacer una prueba con nuestro algoritmo para conocer los empaquetamientos de fragmentos que se alcanzan con un potencial atractivo pero inespecífico. Así, cuando empleemos este método con un potencial hidrófobo, sabremos en qué medida las estructuras que obtengamos son producto del colapso de la estructura o de la afinidad entre pares de residuos específicos.

4.1. Modelo para la proteína y algoritmo de muestreo

Tanto para el test con un potencial uniformemente atractivo como para el estudio de potenciales basados en estructuras, utilizamos proteínas de tipo todo α, formadas fundamentalmente por hélices α. En una proteína, las distintas caras de las hélices interaccionan bien con el disolvente o bien con otras regiones de la proteína, en función de la naturaleza de sus cadenas laterales. Los residuos polares tienen mayor tendencia a quedar expuestos hacia el disolvente, mientras que los hidrófobos tienden a resguardarse del disolvente en el seno de la proteína. Los potenciales hidrófobos que estudiamos en esta parte de la Tesis tratan de reproducir fundamentalmente este tipo de interacciones. Por ello, para nuestro estudio podemos considerar las hélices α como fragmentos rígidos y buscar el mejor empaquetamiento de fragmentos de acuerdo con el potencial.

En la Figura 4.1 mostramos el conjunto de proteínas que hemos seleccionado para el estudio de estos potenciales. Para cada una de las proteínas, tomamos del Protein Data Bank las coordenadas de los átomos que se utilizan como centros de interacción en los distintos potenciales. En algunos casos se consideran sólo los carbonos-α, y en otros, los carbonos-α y los centros de las cadenas laterales. Para obtener las posición del centro de la cadena lateral de un residuo, promediamos las posiciones de sus átomos pesados. Para cada proteína, consideramos tantos fragmentos peptídicos como hélices α hay en la estructura tridimensional de la proteína. Para dividir la secuencia en fragmentos seguimos la orientación que dan el encabezamiento del archivo PDB y el programa STRIDE[124] de asignación de estructura secundaria.

Hemos realizado una serie de cálculos preliminares obviando aquellos residuos que no forman parte de las hélices α, sino que se encuentran en lazos entre hélices. Hemos observado que los residuos de los lazos son decisivos para que el mínimo energético pueda corresponder a la estructura nativa. Probablemente, si es necesario considerar estos

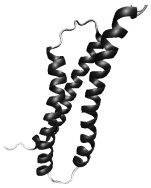
Descripción	Código PDB	nº fragmentos	nº residuos en el modelo	
Proteína ROP	**1rpo**	2	58	
Ovillo de 3 hélices A3D *de novo*	**2a3d**	3	69	
Chaperona reguladora de la familia BAG	**1i6z**	3	116	
Dominio FAT de la quinasa de adhesión focal	**1ktm**	4	128	
Apolipoproteína E4	**1le4**	5	138	
Apolipoporina III	**1ls4**	5	155	

Figura 4.1: Conjunto de proteínas para el estudio de potenciales hidrófobos con el número de fragmentos en el modelo, el número de residuos y una representación de su estructura tridimensional.

residuos no es por sus interacciones específicas, sino por la restricción de volumen excluido que imponen. Así, en nuestro estudio hemos incluido los residuos de los lazos entre hélices α en los fragmentos rígidos. Aquellos residuos que no forman parte de ninguna hélice se incluyen en un fragmento peptídico vecino. La división entre un fragmento y el siguiente corresponde al punto en que cambia la dirección de propagación en la estructura tridimensional de la proteína. La separación entre fragmentos sucesivos es la distancia de un enlace virtual, es decir, la distancia entre dos carbonos-α de residuos que están unidos por enlace peptídico *trans*, que es de 3.8 Å.

La inclusión de los lazos en los fragmentos rígidos tiene influencia en la implementación del algoritmo evolutivo. Como dijimos en los Capítulos 2 y 3, tenemos que utilizar la codificación que resulte más apropiada para el experimento que queramos hacer. En este caso, utilizamos la codificación interna del algoritmo evolutivo, más adecuada para fragmentos conectados entre sí por lazos cortos. En esta codificación, las tres primeras variables de cada grupo (r, θ, φ) definen la posición del primer centro de interacción de un fragmento con respecto al último del fragmento anterior. Como acabamos de comentar, la separación entre fragmentos en este caso es de 3.8 Å, por lo que la primera de las variables tiene un valor fijo. Por tanto, para una proteína de N fragmentos, el método evolutivo tiene que optimizar $5 \times (N-1)$ variables, en lugar de $6 \times (N-1)$.

Los parámetros del algoritmo evolutivo que utilizamos en estos experimentos son los mismos que mostrábamos en la Tabla 3.1. El tamaño de la población es n_c=100 individuos. Cada una de las 10 optimizaciones independientes realizadas con el algoritmo genético evoluciona durante n_{gener}=500 generaciones. La probabilidad de entrecruzamiento de dos puntos p_{cross} es 0.5, igual a la del entrecruzamiento de un solo punto $(1 - p_{cross})$, y la probabilidad de mutación es p_{mut}=0.1. También en este caso, el umbral de $RMSD$ entre las conformaciones cuando se realiza la selección es de 1 Å.

4.2. Minimización con un potencial de colapso inespecífico

En nuestro método, el muestreo conformacional se realiza sobre los posibles empaquetamientos de fragmentos rígidos. Como acabamos de comentar, hemos modificado la definición de fragmentos rígidos con respecto a la puesta a punto del método del Capítulo 3. Ahora, la distancia entre dos fragmentos sucesivos es fija. Esto supone una reducción del número de grados de libertad de la proteína en el modelo, por lo que ahora no son tantas las posibilidades de empaquetar los fragmentos rígidos. Para comprobar que el algoritmo es capaz de acceder a empaquetamientos de fragmentos diferentes del nativo, realizamos una prueba en la que la función de mérito es un potencial uniformemente atractivo. Así, en el estudio de los potenciales basados en estructuras, será verdaderamente relevante que se defina el mínimo en la conformación nativa, en lugar de en cualquiera de las otras estructuras compactas accesibles. Además, esto nos permite separar, en estos potenciales, la componente de colapso inespecífico de la de secuencia.

Una magnitud cuya minimización permite dirigir el muestreo hacia estados más compactos de la proteína es el radio de giro. De hecho, esta magnitud se ha utilizado anteriormente como "coordenada de reacción" en la simulación del plegamiento[125]. El radio de giro (R_g) se define como la distancia cuadrática media del conjunto de átomos de la proteína con respecto a su centro de gravedad[126]

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}. \tag{4.3}$$

En esta ecuación, d_i es la distancia entre el átomo i y el centro de gravedad, y N es el número de centros que se utilizan para el cálculo del radio de giro. Para evitar que en la minimización se superpongan los fragmentos rígidos, hemos introducido una

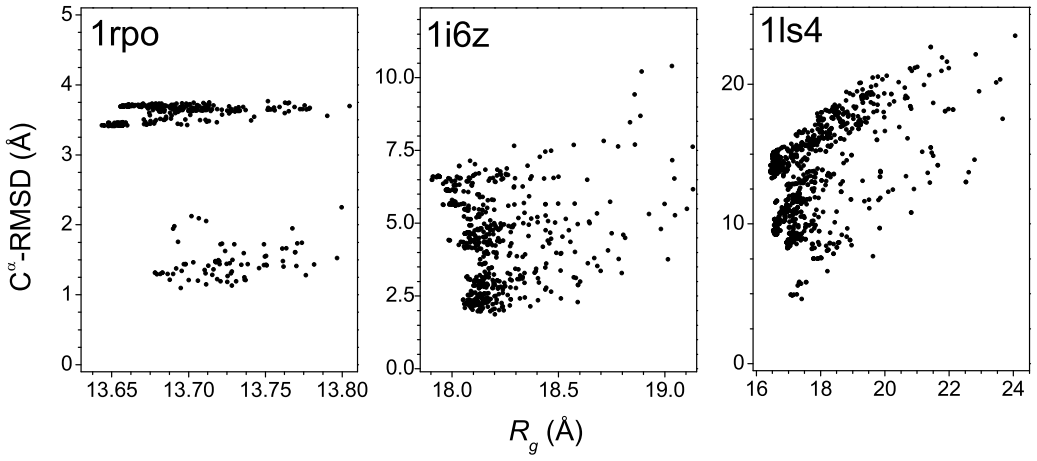


Figura 4.2: Representación del radio de giro frente a $RMSD$ con respecto a la nativa de las mejores conformaciones del las proteínas 1rpo, 1i6z y 1ls4, en la minimización con un potencial de colapso inespecífico.

contribución de volumen excluido entre los centros de interacción.

Hemos llevado a cabo experimentos de minimización del radio de giro para tres de las proteínas de la Figura 4.1, con las que después evaluaremos los potenciales. Hemos escogido tres proteínas de distinto tamaño: 1rpo, de dos fragmentos, 1i6z de tres y 1ls4 de cinco fragmentos. Repetimos cinco veces cada uno de los experimentos de minimización, para los que utilizamos el conjunto de parámetros que hemos comentado en la sección anterior.

En la Figura 4.2 mostramos una representación de los resultados correspondientes a estas minimizaciones. En cada gráfica, representamos el radio de giro frente al valor de $RMSD$ con respecto a la conformación nativa para aquellas conformaciones que han sido las de menor radio de giro a lo largo de la optimización. Este tipo de representación permite que nos hagamos una idea de cómo durante la minimización se explora el espacio conformacional de la proteína con nuestra aproximación de fragmentos rígidos. Para

las tres proteínas aparecen varios mínimos para el radio de giro. Uno de estos mínimos, el de menor $RMSD$, corresponde a una conformación parecida a la nativa. Los restantes mínimos corresponden a conformaciones más o menos diferentes de la nativa, con ordenamientos alternativos de los fragmentos rígidos. Por tanto, la conclusión general de estos resultados es que, a pesar de haber reducido el número de grados de libertad sobre los que se lleva a cabo el muestreo, éste sigue permitiendo acceder a varios empaquetamientos densos de los fragmentos. Dado que un potencial de colapso inespecífico no dirige hacia el empaquetamiento nativo, tiene que ser la componente de secuencia de un potencial de interacción la encargada de favorecer el empaquetamiento nativo frente a los demás.

Hemos estudiado cómo son las conformaciones alternativas de estas tres proteínas a las que hemos accedido con el potencial de colapso. El caso más sencillo es el de la proteína 1rpo, de tan solo dos fragmentos. En este caso se definen varios mínimos, como se puede ver en el panel correspondiente de la Figura 4.2. Representamos varias conformaciones de esta proteína en la Figura 4.3 a nivel de carbonos-α y en diagramas topológicos. En los diagramas topológicos, cada uno de los fragmentos que empaqueta el algoritmo evolutivo se representa como un cilindro con una tonalidad de gris diferente. En la Figura 4.3 (a) representamos la conformación nativa. Uno de los mínimos del radio de giro que hemos obtenido es muy parecido a la conformación nativa. Mostramos esta conformación en la Figura 4.3 (b). Tiene un valor de $RMSD$ con respecto a la nativa de cerca de 1 Å. Con valores de $RMSD$ muy superiores encontramos otro mínimo más profundo y mejor definido. Su estructura la mostramos en la Figura 4.3 (c), donde se ve claramente la distinta orientación de los fragmentos.

La siguiente proteína con la que llevamos a cabo la minimización del radio de giro es la proteína 1i6z, en cuya conformación nativa aparecen tres hélices α. Como se trata de una proteína con un mayor número de fragmentos, hay más posibilidades para

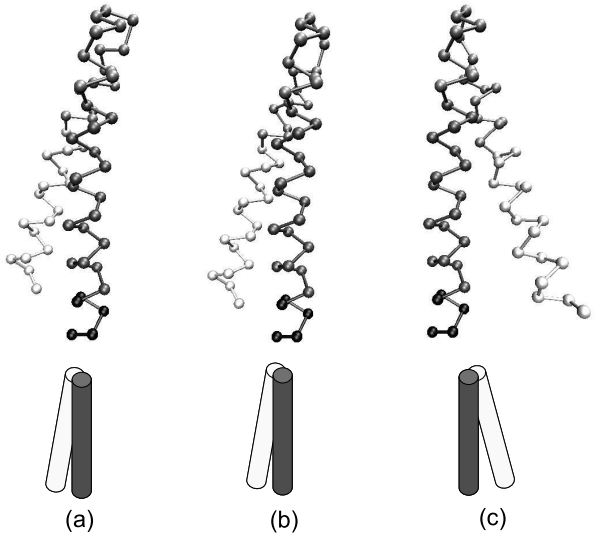


Figura 4.3: Representación del esqueleto y diagramas topológicos de varias conformaciones de la proteína 1rpo obtenidas en la minimización del radio de giro. (a) Conformación nativa, (b) mínimo con $RMSD$=1.0 Å, y (c) mínimo con $RMSD$=3.6 Å.

empaquetarlos. En la representación del radio de giro frente al valor de $RMSD$ de las conformaciones más compactas que se han ido obteniendo a lo largo de la minimización (Figura 4.2), vemos que aparecen hasta tres mínimos diferentes. Mostramos una serie de representaciones de estas conformaciones de la proteína en la Figura 4.4. El mínimo del radio de giro más semejante a la conformación nativa (a) aparece con un valor de $RMSD$ ligeramente inferior a 2.5 Å. A pesar de un leve reordenamiento con respecto a la conformación nativa (ver Figura 4.4 (b)), se trata de una conformación muy parecida. El segundo mínimo que encontramos tiene un valor de $RMSD$ con respecto a la nativa próximo a 4.5 Å (Figura 4.4 (c)). En este caso, la primera hélice (en gris oscuro en la Figura) está enfrentada con las otras dos con una "cara" diferente a la "cara" con la que interacciona en la conformación nativa. La tercera de las conformaciones que mostramos corresponde al tercer mínimo, con valor de $RMSD$ cercano a 6.5 Å (Figura 4.4 (d)). En este caso las hélices están organizadas de una manera distinta que en la nativa. Si imaginamos un eje en torno al cual se dispongan las hélices, en la nativa estarían

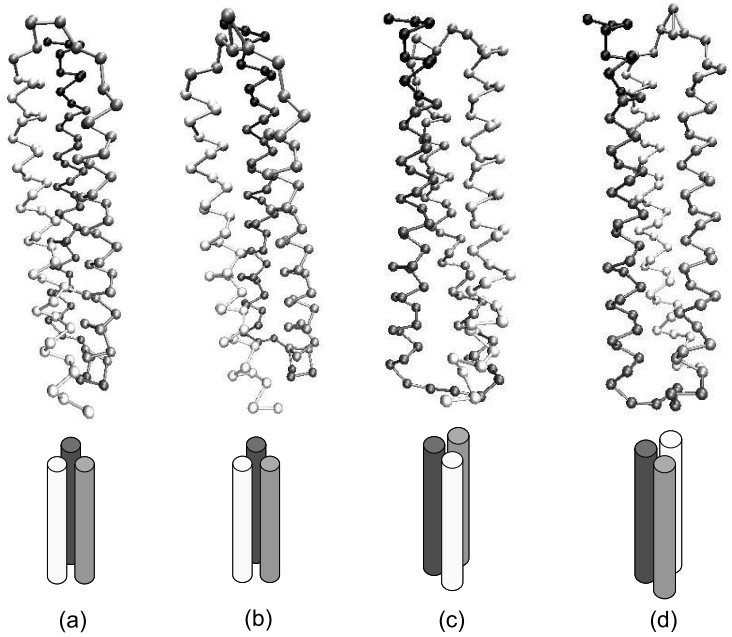


Figura 4.4: Representación del esqueleto y diagramas topológicos de varias conformaciones de la proteína 1i6z obtenidas en la minimización del radio de giro. (a) Conformación nativa, (b) mínimo con $RMSD$=2.3 Å, (c) mínimo con $RMSD$=4.5 Å, y (d) mínimo con $RMSD$=6.4 Å.

dispuestos en el sentido de las agujas del reloj. En cambio, en la conformación con $RMSD \simeq 6.5$ Å se dispondrían en sentido contrario.

Finalmente, en el caso de 1ls4, una proteína en cuya conformación nativa aparecen cinco hélices α, encontramos una situación semejante a la de 1i6z. También en este caso encontramos diferentes mínimos en la superficie explorada por el algoritmo (ver Figura 4.2). En la Figura 4.4 mostramos algunas de estas conformaciones. La más parecida a la nativa (a) tiene un $RMSD$ de cerca de 5 Å. En la Figura 4.4 (b) mostramos representaciones de este mínimo. A pesar de ciertos ajustes de distancias y orientaciones, su topología es semejante a la de la conformación nativa. Mostramos además otros mínimos locales para el radio de giro, que tienen valores de $RMSD$ próximos a 10 y 15 Å, y corresponden a ordenamientos diferentes de los fragmentos rígidos ((c) y (d)).

Con estos experimentos comprobamos que el algoritmo es capaz de alcanzar dis-

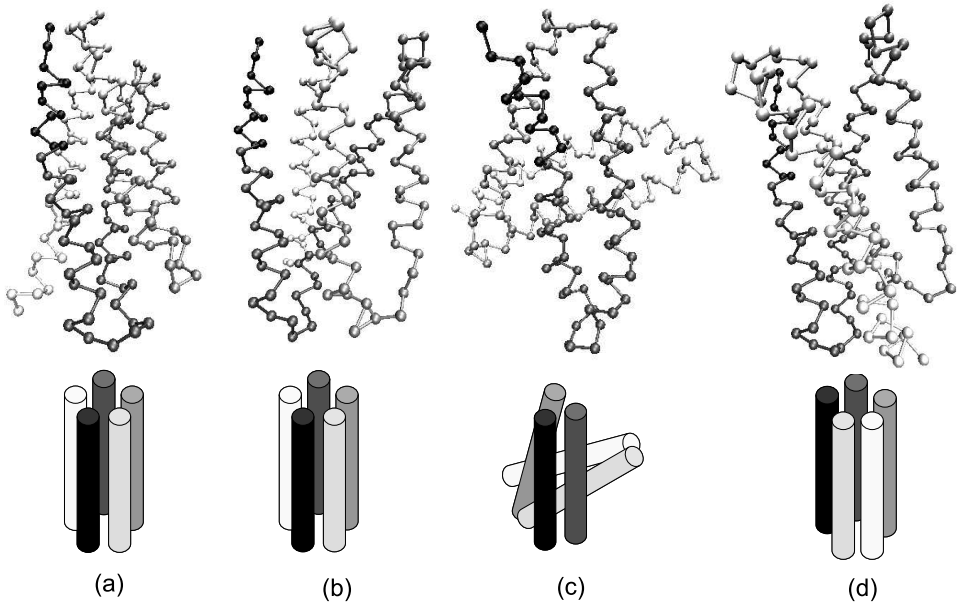


Figura 4.5: Representación del esqueleto y diagramas topológicos de varias conformaciones de la proteína 1ls4 obtenidas en la minimización del radio de giro. (a) Conformación nativa, (b) mínimo con $RMSD$=4.9 Å, (c) mínimo con $RMSD$=14.2 Å, y (d) mínimo con $RMSD$=9.2 Å.

tintas conformaciones densas a pesar de haber restringido la búsqueda con un modelo con sólo cinco grados de libertad por fragmento. De este modo, podemos ver que en ausencia de efectos de secuencia hay distintas soluciones compactas posibles. En los experimentos con los potenciales hidrófobos, que introducimos a continuación, el modelo para las interacciones es el responsable de dirigir el muestreo hacia el empaquetamiento más parecido al nativo.

4.3. Potenciales de interacción hidrófobos

A continuación describimos los tres potenciales de interacción que consideramos en esta parte de nuestro estudio. Todos ellos son potenciales de interacción hidrófobos basados en estructuras, definidos entre pares de aminoácidos. Hemos escogido estos tres potenciales porque son representativos de las distintas aproximaciones que permiten extraer

términos energéticos a partir de estructuras de proteínas. Por otra parte, difieren también en el tipo de dependencia con la distancia y en los centros de interacción que se utilizan para el cálculo de la energía. Finalmente, estos potenciales vienen respaldados por estudios comparativos en los que se ha comprobado su buen comportamiento. Una visión más completa de cada uno de ellos puede encontrarse en los artículos citados.

4.3.1. Potencial de Nanias

El elemento principal del llamado potencial de Nanias es la matriz de energías de contacto de Miyazawa y Jernigan[88], al que Nanias *et al.* añaden una forma funcional sencilla para la dependencia con la distancia entre centros de interacción[100]. Con este potencial, realizaron un estudio parecido al nuestro. Para un grupo numeroso de proteínas todo α, empaquetaban hélices tomadas como fragmentos rígidos para tratar de recuperar la conformación nativa[100]. Nanias *et al.* utilizaron un método de optimización global para su estudio, y las hélices α que empaquetaron, al contrario de como lo hacemos nosotros, tenían geometría ideal. Los buenos resultados de este método nos animaron a utilizar su potencial de interacción. Esto nos permite comparar este potencial sencillo, con una matriz de contactos para las interacciones entre pares y un solo parámetro para la dependencia con la distancia, con otros potenciales que en su diseño tienen una dependencia con la distancia más detallada.

Como hemos comentado, el potencial que utilizan Nanias *et al.*[100] es la matriz de energías de Miyazawa y Jernigan, de dimensión (20×20)[88], que mostramos en la Figura 4.6. Esta matriz de energías quizás sea el mejor establecido y más citado entre los potenciales de campo medio obtenidos para proteínas. Así, en un estudio reciente[127], se compararon 29 potenciales basados en estructuras, entre los que se incluian varias versiones del potencial de Miyazawa y Jernigan[80,88,128]. La conclusión de este estudio fue que la mayoría de los potenciales están correlacionados con alguna de las matrices

de contactos de Miyazawa y Jernigan. Los términos de la matriz que utilizan Nanias *et al.*[100] son las energías efectivas de contacto entre pares de aminoácidos e_{ij} (ver Figura 4.6). Para obtener estas energías, Miyazawa y Jernigan utilizan la aproximación cuasiquímica y un tratamiento aproximado de los efectos de la conectividad de las cadenas polipeptídicas. El estado de referencia para el cálculo del potencial de campo medio es una mezcla aleatoria de aminoácidos no unidos y moléculas de disolvente. Los centros de interacción que se consideran en el potencial son los centros de las cadenas laterales. Se considera que hay un contacto entre aquellos residuos cuyos centros están separados por una distancia menor o igual que 6.5 Å.

Por tratarse de un potencial de contacto, para un par de residuos de tipos i y j, el potencial de Miyazawa y Jernigan tiene la forma de un pozo cuadrado. Esta dependencia con la distancia tan abrupta hace que su aplicación se limite a estudios realizados con métodos que utilicen estructuras estáticas, como el enhebrado, o bien a simulaciones en red. Para su estudio en el espacio continuo, Nanias *et al.* introducen una función sencilla para la dependencia con la distancia. Se trata de un potencial de tipo Lennard-Jones, que permite calcular la energía de un término entre pares según la ecuación

$$u(r_{ij}) = \frac{e'_{ij}}{14 \pm 15}\left[14\left(\frac{\sigma}{r_{ij}}\right)^{15} \pm 15\left(\frac{\sigma}{r_{ij}}\right)^{14}\right]. \qquad (4.4)$$

En esta ecuación, e'_{ij} es la energía de alineamiento, que se define como la resta $e_{ij} - e_{rr}$, y es el valor adecuado para calcular las energías[88]. Los signos de la Ecuación 4.4 se ajustan de acuerdo con el valor de e'_{ij}: positivo si $e'_{ij}>0$ y negativo si $e'_{ij}<0$. Nanias *et al.* utilizan los carbonos-α como centros de interacción y r_{ij} es la distancia entre los carbonos-α de los residuos i y j. Por tanto, no utilizan los mismos centros de interacción que Miyazawa y Jernigan consideraron para diseñar el potencial, los centros de las cadenas laterales. Por este motivo, a la distancia σ se le asigna el valor de 7.5 Å en lugar de 6.5 Å.

Table 3. Contact energies in RT units; e_{ij} for upper half and diagonal and $\acute{e}_{ij}$ for lower half

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
Cys	−5.44	−4.99	−5.80	−5.50	−5.83	−4.96	−4.95	−4.16	−3.57	−3.16	−3.11	−2.86	−2.59	−2.85	−2.41	−2.27	−3.60	−2.57	−1.95	−3.07	Cys
Met	−5.46	−5.46	−6.56	−6.02	−6.41	−5.32	−5.55	−4.91	−3.94	−3.39	−3.51	−3.03	−2.95	−3.30	−2.57	−2.89	−3.98	−3.12	−2.48	−3.45	Met
Phe	−0.20	−0.20	−7.26	−6.84	−7.28	−6.29	−6.16	−5.66	−4.81	−4.13	−4.28	−4.02	−3.75	−4.10	−3.48	−3.56	−4.77	−3.98	−3.36	−4.25	Phe
Ile	−0.01	−0.01	0.06	−6.54	−7.04	−6.05	−5.78	−5.25	−4.58	−3.78	−4.03	−3.52	−3.24	−3.67	−3.17	−3.27	−4.14	−3.63	−3.01	−3.76	Ile
Leu	0.01	0.03	−0.08	−7.37	−6.48	−6.14	−5.67	−4.91	−4.16	−4.34	−3.92	−3.74	−4.04	−3.40	−3.59	−4.54	−4.03	−3.37	−4.20	Leu	
Val	0.18	0.10	−0.01	−0.04	−5.52	−5.18	−4.62	−4.04	−3.38	−3.46	−3.05	−2.83	−3.07	−2.48	−2.67	−3.58	−3.07	−2.49	−3.32	Val	
Trp	−0.29	0.00	0.02	0.08	0.11	−5.06	−4.66	−4.17	−3.36	−3.42	−3.22	−2.99	−3.07	−3.11	−2.84	−2.99	−3.98	−3.41	−2.69	−3.73	Trp
Tyr	−0.10	0.05	0.11	0.10	0.23	−0.04	−4.17	−3.36	−3.01	−3.01	−2.78	−2.76	−2.97	−2.76	−2.79	−3.52	−3.16	−2.60	−3.19	Tyr	
Ala	0.15	0.17	0.05	0.13	0.08	0.07	0.09	−2.72	−2.31	−2.32	−2.01	−1.84	−1.89	−1.70	−1.51	−2.41	−1.83	−1.31	−2.03	Ala	
Gly	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	−2.24	−2.08	−1.82	−1.74	−1.66	−1.59	−1.22	−2.15	−1.72	−1.15	−1.87	Gly	
Thr	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	−2.12	−1.96	−1.88	−1.90	−1.80	−1.74	−2.42	−1.90	−1.31	−1.90	Thr	
Ser	0.53	0.44	0.59	0.60	0.55	0.38	0.14	0.18	0.14	−0.06	−1.67	−1.58	−1.49	−1.63	−1.48	−2.11	−1.62	−1.05	−1.57	Ser	
Asn	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	−1.68	−1.71	−1.68	−1.51	−2.08	−1.64	−1.21	−1.53	Asn	
Gln	0.20	0.30	0.37	0.42	0.46	0.19	−0.12	0.24	0.24	−0.08	0.11	−0.10	−1.54	−1.46	−1.42	−1.98	−1.80	−1.29	−1.73	Gln	
Asp	0.77	0.75	0.71	0.89	0.89	0.30	−0.07	0.26	0.13	−0.14	−0.19	−0.24	−0.09	−1.21	−1.02	−2.32	−2.29	−1.68	−1.33	Asp	
Glu	0.30	0.52	0.46	0.55	0.55	0.00	−0.25	0.30	0.36	−0.22	−0.19	−0.21	−0.19	0.05	−0.91	−2.15	−2.27	−1.80	−1.26	Glu	
His	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.50	0.16	0.26	0.29	0.31	−0.19	−0.16	−3.05	−2.16	−1.35	−2.25	His	
Arg	0.38	0.42	0.41	0.43	0.47	−0.11	−0.30	0.30	0.18	−0.07	−0.01	−0.02	−0.26	−0.91	−1.04	0.14	−1.55	−0.59	−1.70	Arg	
Lys	0.31	0.33	0.32	0.37	0.33	−0.10	−0.46	0.11	0.03	−0.19	−0.15	−0.30	−0.46	−1.01	−1.28	0.23	0.24	−0.12	−0.97	Lys	
Pro	0.16	0.25	0.39	0.35	0.31	−0.33	−0.23	0.20	0.13	0.04	0.14	0.18	−0.08	0.14	0.07	0.15	−0.05	−0.04	−1.75	Pro	
$e_r - 2.55$	e_r	−3.57	−3.92	−4.76	−4.42	−4.81	−3.89	−3.81	−3.41	−2.57	−2.19	−2.29	−1.98	−1.92	−2.00	−1.84	−1.79	−2.56	−2.11	−1.52	−2.09
$e_t - 3.60$	e_t	−4.29	−4.73	−5.57	−5.29	−5.71	−4.72	−4.41	−3.87	−3.17	−2.53	−2.63	−2.27	−2.14	−2.35	−2.02	−2.07	−2.94	−2.43	−1.82	−2.53
$f_i - 3.60$	f_i	−5.58	−6.14	−7.39	−7.09	−7.88	−6.15	−5.34	−4.60	−3.24	−2.22	−2.48	−1.92	−1.74	−1.93	−1.54	−1.49	−2.91	−2.07	−1.17	−1.97
N_r/N_i	2.096	2.722	2.780	2.811	2.893	2.728	2.537	2.493	2.143	1.840	1.973	1.771	1.699	1.720	1.598	1.508	2.075	1.787	1.343	1.629	
q_i 7.162	6.281	6.646	6.137	5.870	6.042	6.087	6.155	5.793	6.037	6.334	6.284	6.486	6.582	6.574	6.469	6.487	6.235	6.241	6.318	6.569	5.858

Figura 4.6: Matriz de energías de contacto de Miyazawa y Jernigan[88] utilizada para el cálculo de la energía en el potencial de Nanias.

La energía de una conformación se obtiene como la suma de todos los términos $u(r_{ij})$ entre residuos que pertenecen a diferentes fragmentos peptídicos, cuyos carbonos-α se encuentren a una distancia inferior a una distancia de corte de $2.5\times\sigma$. Como en el Capítulo 3, no calculamos la energía entre residuos del mismo fragmento porque permanece constante en todas las conformaciones. Tampoco se contabilizan las interacciones entre residuos de distintos fragmentos en el modelo separados por menos de 4 residuos en la secuencia.

4.3.2. Potencial TE-13

Como hicieron antes otros autores[84], Dror Tobi y Ron Elber obtienen el potencial TE-13 empleando una aproximación matemática, no basada en la física[92,93]. Su punto de partida para generar el potencial es la hipótesis termodinámica[4], es decir, que para una proteína la energía tiene su mínimo en la conformación nativa. Para obtener el potencial generan muchas estructuras alternativas o quimeras para muchas secuencias. En este caso las quimeras se generan por enhebrado de secuencias y utilizando MONSSTER[129], un programa de predicción de estructura. Para cada secuencia, se impone la condición de que la energía calculada con el conjunto de parámetros sea superior a la de la nativa, como hemos expresado en la Ecuación 4.2. Para obtener sus términos de energía, tratan de resolver por programación lineal un conjunto de hasta 32.442.176 inecuaciones[93]. El mejor resultado que Tobi y Elber obtienen es el potencial entre pares denominado TE-13, que tiene una forma funcional muy libre. Para cada par de residuos, el potencial se define en un intervalo de distancias entre 2 y 9 Å entre centros de cadenas laterales. Este intervalo, a su vez, se divide en 13 segmentos. Por tanto, podemos pensar en el potencial de Tobi y Elber como en 13 matrices 20×20 como la de Miyazawa y Jernigan, una para cada segmento dentro del intervalo entre 2 y 9 Å.

En sus trabajos, Tobi y Elber comparan su potencial con potenciales de contac-

to[88,130–132] y con potenciales que tienen dependencia detallada con la distancia[133]. La condición que imponen los autores consiste en que, para un grupo muy numeroso de proteínas, un potencial sea capaz de reconocer la correspondiente conformación nativa para un conjunto de secuencias. Este reconocimiento consiste, una vez más, en que a la conformación nativa se le asigne la mínima energía. Tobi y Elber muestran en sus resultados que el potencial tiene una eficacia semejante a la de los demás potenciales, pero que es más uniforme y consistente[93].

Para calcular la energía de una conformación de una proteína con este potencial, se suman las contribuciones entre pares de residuos que en nuestro modelo están en fragmentos diferentes. Para cada par de residuos, se determina la distancia entre centros de cadenas laterales, y se le asigna la energía correspondiente al tipo de residuos y al segmento en el que se encuentra esa distancia. Hemos observado que con esta manera de calcular la energía, en nuestro muestreo conformacional se estabilizan estructuras en las que hay átomos solapando. Para resolver este problema, añadimos dos contribuciones de volumen excluido al potencial: uno para centros de cadenas laterales, a distancias inferiores a 2 Å, donde el potencial no está definido, y otro entre pares de carbonos-α a distancias inferiores a 3.8 Å. Como con el potencial de Nanias, no consideramos interacciones entre pares de residuos que estén separados por menos de 4 residuos en la secuencia.

4.3.3. Potencial DFIRE-SCM

El tercero y más reciente de los potenciales que incluimos en esta evaluación es el llamado potencial DFIRE-SCM. Este potencial es uno de los tres que ha obtenido el grupo de Zhou[96,123] con una misma estrategia. Los potenciales de la familia DFIRE son potenciales estadísticos, basados como el de Miyazawa y Jernigan en la Ecuación 4.1. Por tanto, también Zhou *et al.* obtienen sus valores para la energía a partir del número de contactos

entre tipos de aminoácidos que aparecen en estructuras cristalográficas de un grupo de proteínas. La particularidad de los potenciales DFIRE reside en el estado de referencia que utilizan los autores, que se asemeja a un "gas ideal finito". Para definir este estado de referencia se genera una distribución uniforme de esferas no interaccionantes. La tendencia a interaccionar de dos residuos se calcula con respecto a esta distribución. Otra característica del potencial DFIRE-SCM es una definición detallada de la dependencia con la distancia entre centros de interacción. Como en el caso del potencial TE-13, los centros de interacción son los centros de las cadenas laterales. En este caso, el intervalo de distancia entre centros, entre 2 y 15 Å, se divide en 20 segmentos.

Parece que sea este estado de referencia lo que hace que los potenciales de la familia DFIRE mejoren los resultados de otros potenciales. Zhou *et al.* han realizado pruebas de dos tipos. En primer lugar, compararon su potencial atómico con otros dos potenciales[134,135] utilizando quimeras[123]. En otro estudio realizaron experimentos de minimización con una búsqueda restringida en el espacio de conformaciones de la proteína[136]. En ambos trabajos los resultados obtenidos con el potencial fueron muy alentadores.

En nuestro estudio, para calcular la energía de una conformación de la proteína, se suman las energías entre pares de residuos que se encuentran en fragmentos diferentes de la proteína. Para un par de residuos, se calcula la distancia entre centros de cadenas laterales y se asigna el valor de energía correspondiente al tipo de residuos de que se trate. Como en el caso del potencial de Tobi y Elber, añadimos dos contribuciones de volumen excluido: una entre carbonos-α (para distancias menores de 3.8 Å) y otra entre centros de cadenas laterales (para distancias menores de 2 Å). Los propios autores del potencial, como nosotros, han añadido en sus minimizaciones una repulsión en el cálculo de la energía[136]. Como hemos indicado para los otros potenciales, tampoco se contabilizan las interacciones entre residuos de la cadena separados por menos de 4

residuos en la secuencia.

4.4. Minimización de la energía con potenciales hidrófobos

Con los tres potenciales que acabamos de describir, hemos llevado a cabo una serie de experimentos de minimización con nuestra estrategia evolutiva para las seis proteínas de la Tabla 4.1. Para cada uno de las potenciales, la función de mérito del algoritmo es la energía de las interacciones entre fragmentos. Los parámetros de la estrategia evolutiva son los mismos que hemos utilizado en la Sección 4.2.

4.4.1. Potencial de Nanias

En la Tabla 4.1 mostramos los resultados que hemos obtenido con el potencial de Nanias para la minimización de la energía. Presentamos el valor que asigna el potencial para la energía de la conformación nativa, la energía de la conformación de mínima energía que se ha alcanzado y el valor de $RMSD$ de esta conformación frente a la nativa en angstroms. El primer aspecto que queremos destacar en estos resultados es el valor que el potencial de Nanias asigna a la conformación nativa. Para todas y cada una de las proteínas del conjunto estudiado, la energía es marcadamente positiva. Como los propios Nanias *et al.* nos comentaron en comunicación personal, estos valores no son muy significativos. Se deben a que la función energética de tipo Lennard-Jones tiene una región atractiva muy estrecha, y está controlada por un solo parámetro, la distancia de equilibrio σ. Probablemente, los autores buscaron un valor para este parámetro que se adaptase al intervalo de distancias entre centros habitual en estructuras de proteínas. Sin embargo, no consideraron la variedad en el tamaño de las cadenas laterales que influye

$PDBid$	$n^o\ frag$	$E_{Nat}\ (RT)$	$E_{Min}\ (RT)$	$RMSD$ (Å)
1rpo	2	6239	-39.9	3.7
2a3d	3	1655	-76.4	3.0
1i6z	3	50141	-106.8	2.0
1ktm	4	16463	-108.5	11.0
1le4	5	20411	-158.3	4.2
1ls4	5	41889	-127.9	24.0

Tabla 4.1: Resultados de la minimización con el potencial de Nanias: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}) y $RMSD$ del mínimo con respecto a la nativa. Los valores de energía se expresan en unidades de RT y los de $RMSD$ en amstrongs.

en el empaquetamiento. Por este motivo es normal que para cada una de las proteínas de la Tabla 4.1 encontremos varios residuos para los que ese valor de σ no es válido. En general, se trata de pares de residuos hidrófobos o pares hidrófobo-polar con carga. Estos residuos se encuentran empaquetados muy densamente en la estructura nativa, por lo que la distancia entre sus carbonos-α es pequeña. Al utilizar esta distancia para calcular la energía con la función de tipo Lennard-Jones (ver Ecuación 4.4) se alcanzan valores muy altos.

Como ejemplo de este comportamiento mostramos en la Figura 4.7 el mapa de interacciones de la proteína 1rpo obtenido con el potencial de Nanias. En el mapa se muestran valores de la energía entre pares de residuos en dos conformaciones de la proteína, la nativa y la minimizada. Los valores de energía de la nativa se muestran en el cuadrante (a) y los de la estructura de mínima energía en el (b). Las regiones punteadas del mapa corresponden, para ambas conformaciones, a la superficie de interacción entre las dos hélices α. Como hemos visto en la Tabla 4.1, el valor de energía de la conformación nativa de 1rpo es muy elevado. Así, en el cuadrante (a), que corresponde a esta conformación, vemos que aparece un gran número de repulsiones en la conformación nativa, representadas en negro. Si atendemos a los mismos residuos en el cuadrante (b), que corresponde a la conformación minimizada, no aparece ningún contacto repulsivo.

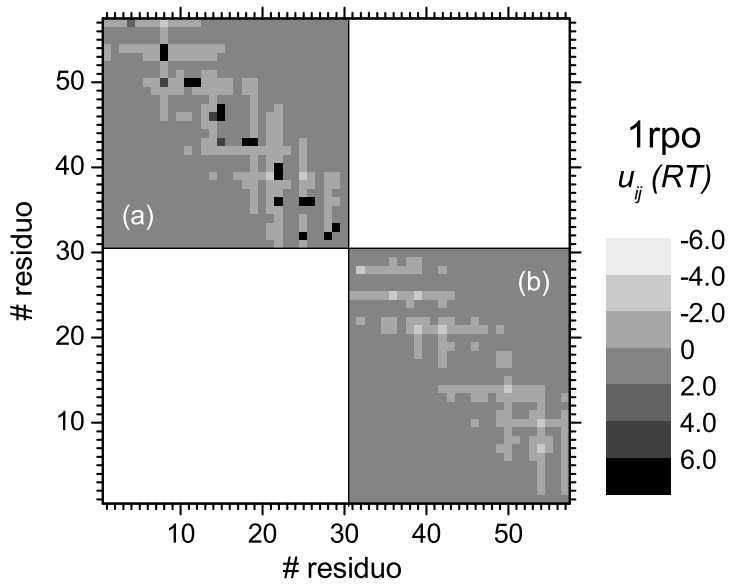


Figura 4.7: Mapa de energías de 1rpo con el potencial de Nanias. Representamos valores de la energía calculada con este potencial para cada par de residuos de distintos fragmentos en las conformaciones (a) nativa y (b) minimizada.

En la optimización, por tanto, se alcanza una conformación en la que los dos fragmentos se encuentran más alejados entre sí que en la conformación nativa. Este alejamiento y un leve cambio de orientación en los fragmentos justifica el elevado valor de $RMSD$ respecto a la nativa (3.7 Å).

En la Figura 4.8 mostramos una representación de la energía frente a $RMSD$ de los mejores individuos en algún punto de la optimización con el potencial de Nanias para varias proteínas. Como comentamos en la Sección 4.2, este tipo de representación nos permite hacernos una idea del muestreo conformacional que el algoritmo realiza sobre la superficie de energía para cada proteína. El panel de la izquierda de la Figura corresponde a la minimización para 1rpo. Si nos fijamos en los caminos que van dibujando las conformaciones representadas, vemos que en la superficie de energía aparecen dos mínimos. La conformación a la que corresponde el mapa de energías mostrado en la Figura 4.7 es la de mayor $RMSD$, es decir, la menos parecida a la nativa. El segundo

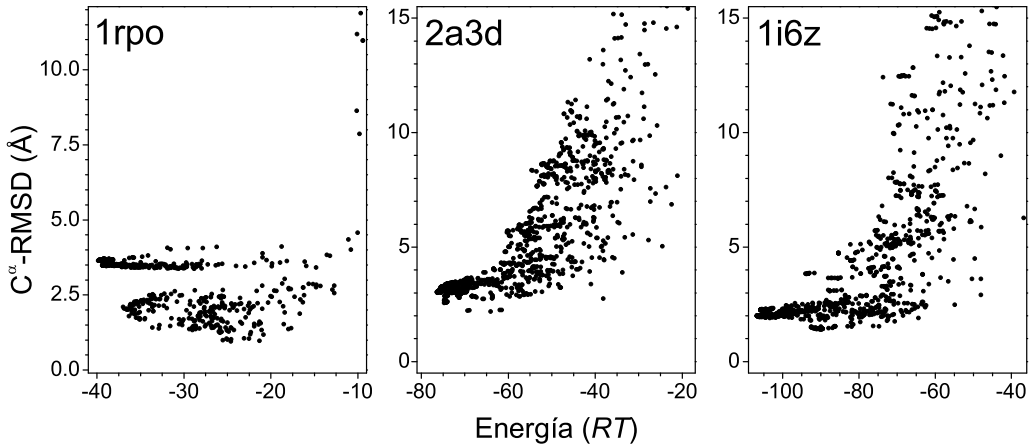


Figura 4.8: Representación de la energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1rpo, 2a3d y 1i6z, en la minimización con el potencial de Nanias.

conjunto de estructuras de baja energía aparece con un valor de $RMSD$ próximo a 2 Å. Debido al volumen excluido que impone la función que definen Nanias *et al.*, para esta conformación más parecida a la nativa aparece un mínimo local. Por tanto, la forma funcional del potencial impide la correcta definición de la superficie de energía.

Este comportamiento que hemos descrito para la proteína 1rpo es generalizable al resto de proteínas con el potencial de Nanias, lo cual supone uno de sus principales problemas. En todos los casos, en la conformación nativa aparece un gran número de repulsiones, por lo que en la minimización se alcanzan conformaciones con energía negativa más o menos alejadas estructuralmente de la nativa. Esto se observa incluso en casos en los que el potencial funciona razonablemente bien, como las proteínas 1i6z y 2a3d, que en el modelo dividimos en tres fragmentos. En ambos casos, las conformaciones de mínima energía tienen la misma topología que la conformación nativa. Por tanto, en estos casos se conservan las buenas características de la matriz de energías de Miyazawa y Jernigan para capturar las interacciones en el seno de la proteína. También para estas dos proteínas mostramos, en la Figura 4.8, la representación de las mejores conformaciones a lo largo de la optimización. En ambos casos la superficie de energía

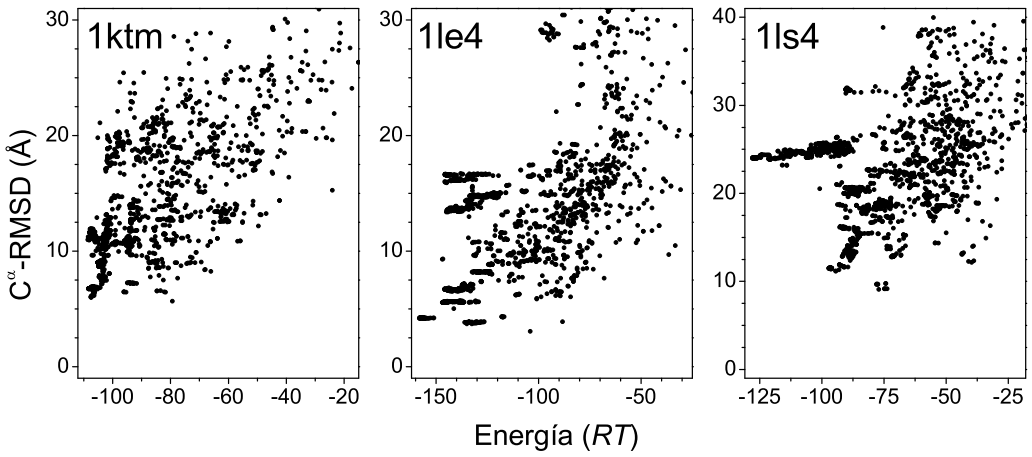


Figura 4.9: Representación de la energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1ktm, 1le4 y 1ls4, en la minimización con el potencial de Nanias.

permite que la minimización transcurra suavemente, sin que el muestreo se quede estancado en mínimos locales. Aun así, en ambos casos los valores de $RMSD$ son altos. La conformación nativa resulta inaccesible por el volumen excluido del potencial.

Para proteínas de mayor tamaño, los resultados de este potencial empeoran notablemente, como puede deducirse de los elevados valores de $RMSD$ del mínimo de energía frente a la estructura nativa (ver Tabla 4.1). Sobre todo para las proteínas 1ktm y 1le4, de cuatro y cinco fragmentos en el modelo, respectivamente, la conformación minimizada es muy distinta de la nativa. En la Figura 4.9 mostramos una representación de los valores de energía y $RMSD$ de las mejores conformaciones a lo largo de la minimización para 1ktm, 1le4 y 1ls4. En el caso de 1ktm, aparecen una serie de conformaciones de baja energía alejadas de la nativa. En concreto, en la conformación de mínima energía, con $RMSD$=11 Å, tres de los cuatro fragmentos están empaquetados formando un haz de hélices, y el cuarto interacciona sólo parcialmente con el lado externo del haz. En el caso de 1ls4, aparece un mínimo global para la energía muy diferente de la conformación nativa (ver Figura 4.9). Esta conformación de mínima energía está formada por dos dominios, uno de dos y otro de tres fragmentos. Estas diferencias con los empaqueta-

mientos nativos se deben de nuevo al volumen excluido del potencial. Para la proteína 1le4 el potencial funciona sustancialmente mejor. En este caso la minimización consigue alcanzar un mínimo cuyo $RMSD$ frente a la estructura nativa es de 4.2 Å. Mostramos la representación de la energía frente a $RMSD$ de las mejores conformaciones de la población a lo largo de la optimización en el panel central de la Figura 4.9. En efecto, vemos un conjunto de conformaciones de muy baja energía para valores de $RMSD$ en torno a 4 Å. Pero, además, hay otra serie de optimizaciones que quedan atrapadas en mínimos locales con valores de $RMSD$ muy superiores y mejor definidos. Este estancamiento del muestreo en mínimos locales se debe a la superficie de energía con grandes barreras que define el potencial.

4.4.2. Potencial TE-13

Los resultados que hemos obtenido con el potencial TE-13 son mucho más satisfactorios que los del potencial de Nanias. En la Tabla 4.2 mostramos valores para la energía de la conformación nativa, la minimizada y el valor de $RMSD$ entre ellas. Los valores de energía obtenidos en este caso están dados en una escala arbitraria y, por tanto, no son comparables a los de Nanias.

Para todas las proteínas del conjunto que estamos utilizando, la energía de la

$PDBid$	n^o $frag$	E_{Nat}	E_{Min}	$RMSD$ (Å)
1rpo	2	-102.3	-155.2	1.0
2a3d	3	-163.5	-260.5	0.9
1i6z	3	-95.2	-296.2	1.2
1ktm	4	-235.7	-439.1	2.8
1le4	5	-362.8	-566.4	1.5
1ls4	5	-163.6	-452.5	18.5

Tabla 4.2: Resultados de la minimización con el potencial TE-13: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}) y $RMSD$ en angstroms del mínimo con respecto a la nativa.

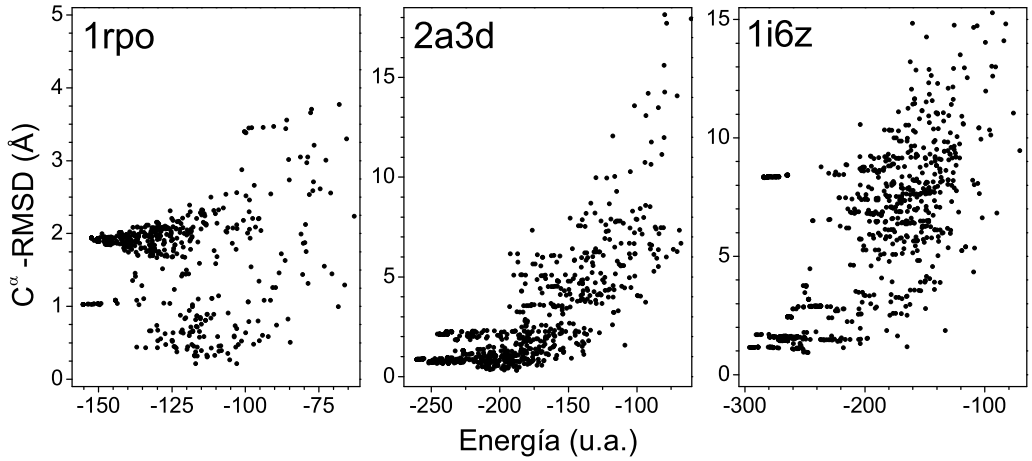


Figura 4.10: Representación de la energía en unidades arbitrarias frente a *RMSD* con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1rpo, 2a3d y 1i6z, en la minimización con el potencial TE-13.

conformación nativa calculada con el potencial es muy inferior en valor absoluto a la del mínimo energético. Esta gran diferencia en valor energético no va ligada a una gran diferencia en estructura. Para todas las proteínas menos 1ls4, el valor de *RMSD* del mínimo frente a la nativa es bajo. Pensamos que estas diferencias en energía se deben a que la forma del potencial de interacción es muy libre[92,93], por haber sido derivado a partir de un ajuste matemático. Para un par de aminoácidos dado, entre un segmento y el siguiente en el intervalo de distancias puede haber una diferencia en el valor de energía de hasta un orden de magnitud. Esto puede verse muy claramente en las representaciones del potencial de Tobi y Elber frente al de Bahar y Jernigan[133], mostrado en la Figura 1 de la referencia de Tobi y Elber (2000)[93]. Esto explica que en la búsqueda conformacional haya un gran salto energético entre la conformación nativa y otra estructura apenas un poco diferente de ella.

En la Figura 4.10 mostramos la representación de las mejores conformaciones a lo largo de la optimización con este potencial para las proteínas 1rpo, 2a3d y 1i6z. Cada una de las conformaciones viene definida por su valor de energía y de *RMSD* frente a

la nativa. En el caso de 1rpo, el algoritmo de minimización localiza dos mínimos en la superficie de energía, uno con $RMSD$ próximo a 1 Å y otro con $RMSD$ de cerca de 2 Å. Estas estructuras son semejantes entre sí. Pero es significativo que la más parecida de ellas a la conformación nativa sea la peor definida en la superficie de energía, como indica la despoblada hilera de puntos en la Figura. Aun así, el potencial es capaz de asignarle una energía menor que al mínimo de mayor $RMSD$.

Para las proteínas de tres fragmentos, 2a3d y 1i6z, el mínimo energético también se define en la conformación nativa de la proteína, como indican los valores de $RMSD$ frente a la nativa, próximos a 1 Å (ver Tabla 4.2). En la Figura 4.10 mostramos la representación de energía y $RMSD$ de las mejores conformaciones a lo largo de la optimización para estas dos proteínas. En el caso de 2a3d, en algunas optimizaciones no se alcanza el mínimo más parecido a la estructura nativa, sino que se quedan atrapadas en otro mínimo local de $RMSD \simeq 2$ Å. En el caso de 1i6z, como en el de 1rpo, aparece un segundo mínimo con energía muy próxima a la del mínimo nativo y $RMSD$ mayor de 8 Å. De nuevo, el mínimo nativo corresponde a un conjunto más o menos disperso de conformaciones en el diagrama. Esto puede implicar una mala definición de la superficie energética. Por su parte, el mínimo con $RMSD \simeq 8$ Å corresponde a un empaquetamiento alternativo de las hélices. En la Figura 4.11 mostramos representaciones del esqueleto y diagramas topológicos para la conformación nativa y los dos mínimos de 1i6z. Podemos describir las conformaciones en función de la disposición de los fragmentos helicoidales en torno a un eje imaginario. En los diagramas topológicos de la Figura 4.11, los fragmentos se representan vistos desde arriba, en un plano perpendicular a ese eje. En la conformación nativa (a) los fragmentos estarían ordenados a favor del sentido de las agujas del reloj, igual que en el mínimo con $RMSD$=1.2 (b). En cambio, en el mínimo alternativo con $RMSD \simeq 8$ Å (c) estarían en contra. Este tipo de "isómero topológico" lo encontramos también en la prueba que hemos realizado con

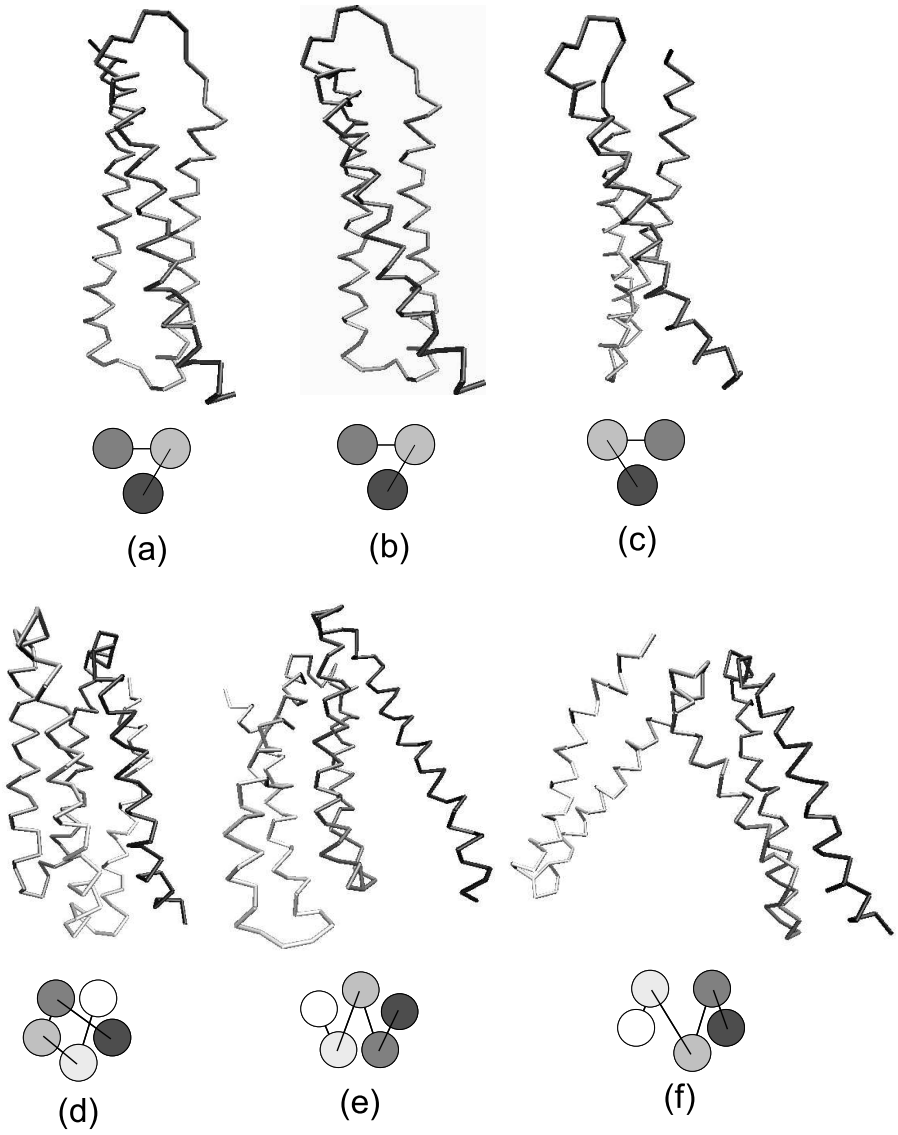


Figura 4.11: Representación del esqueleto y la topología de las proteínas 1i6z y 1ls4 en sus conformaciones nativas y de mínima energía obtenidas con el potencial TE-13. 1i6z: (a) conformación nativa, (b) mínimo con $RMSD$=1.2 Å, (c) mínimo con $RMSD$=8.4 Å. 1ls4: (d) conformación nativa, (e) mínimo con $RMSD$=12.1 Å, (f) mínimo con $RMSD$=18.5 Å.

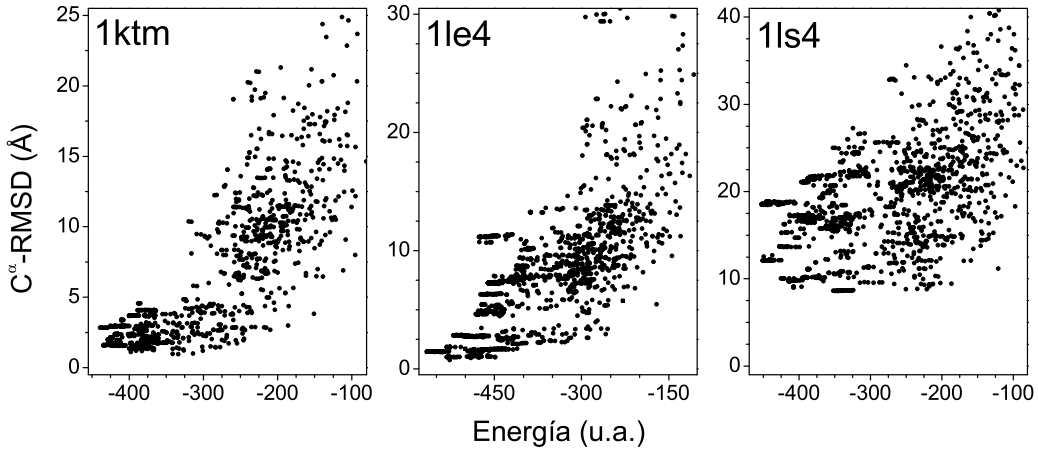


Figura 4.12: Representación de energía en unidades arbitrarias frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1ktm, 1le4 y 1ls4, en la minimización con el potencial TE-13.

un potencial de atracción inespecífico (ver Figura 4.4 (d)).

También para proteínas con mayor número de fragmentos, como 1ktm o 1le4, el mínimo con el potencial TE-13 corresponde a conformaciones con valores de $RMSD$ bajos, es decir, en las vecindades de la conformación nativa (ver Tabla 4.2). Aun así, para ambas proteínas encontramos una cierta dispersión en la definición del mínimo. Esto lo podemos ver en las representaciones de las mejores conformaciones a lo largo de la optimización que mostramos en la Figura 4.12. Por tanto, el comportamiento es parecido al que hemos explicado para proteínas más pequeñas. El mínimo energético aparece, correctamente, en la conformación nativa, pero la superficie de energía se define de una manera relativamente imprecisa, con riesgo de que la búsqueda quede atrapada en mínimos locales.

La proteína para la que peores resultados obtenemos con el potencial TE-13 es 1ls4, de cinco fragmentos en nuestro modelo. En la representación de energía y $RMSD$ frente a la nativa de los mínimos a lo largo de la optimización (ver Figura 4.12) aparecen dos mínimos, aproximadamente con el mismo valor para la energía. Ninguno de ellos corresponde a la conformación nativa. En la Figura 4.11 mostramos las representaciones

esquemáticas y diagramas topológicos de varias conformaciones para esta proteína. En la Figura 4.11 (d) representamos la conformación nativa de 1ls4. Uno de los mínimos energéticos es un haz de hélices con una disposición alternativa de los fragmentos (e). Este tipo de empaquetamiento, en el que los fragmentos se disponen en un orden diferente al de la conformación nativa, lo encontrábamos tambien en nuestra prueba con el potencial uniformemente atractivo. Representamos el otro mínimo, degenerado en energía, en la Figura 4.11 (f). En esta conformación los fragmentos se agrupan en dos dominios, con dos y tres fragmentos rígidos.

4.4.3. Potencial DFIRE-SCM

En la Tabla 4.3 mostramos los resultados para el conjunto de proteínas obtenidos con el potencial DFIRE-SCM. En este caso, la diferencia entre energía de la conformación de mínima energía y la conformación nativa es relativamente pequeña, al contrario de lo que sucedía con el potencial TE-13. Esta diferencia en energía es coherente con la pequeña diferencia entre la estructura del mínimo energético y la conformación nativa. En la Tabla vemos que para todas las proteínas que hemos estudiado, el valor de $RMSD$ es próximo, y a veces inferior, a 1 Å. Estos valores de $RMSD$ tan bajos reflejan el excelente funcionamiento que tiene el potencial en nuestros experimentos de minimización

$PDBid$	n^o $frag$	E_{Nat} (Kcal/mol)	E_{Min} (Kcal/mol)	$RMSD$ (Å)
1rpo	2	-23.0	-24.2	0.4
2a3d	3	-33.4	-42.5	0.7
1i6z	3	-52.6	-65.4	1.0
1ktm	4	-89.1	-104.9	0.5
1le4	5	-101.8	-112.0	1.0
1ls4	5	-127.9	-128.2	1.2

Tabla 4.3: Resultados de la minimización con el potencial DFIRE-SCM: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}) y $RMSD$ en angstroms del mínimo con respecto a la nativa.

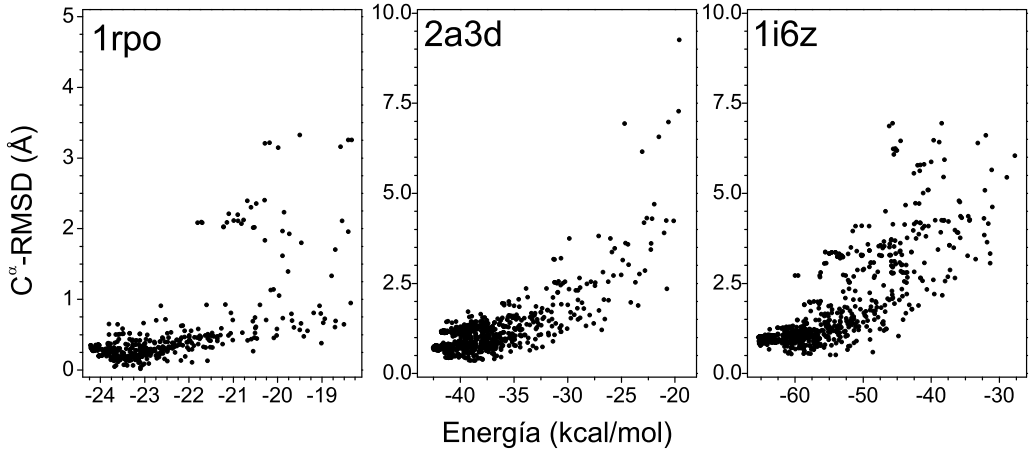


Figura 4.13: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1rpo, 2a3d y 1i6z, en la minimización con el potencial DFIRE-SCM.

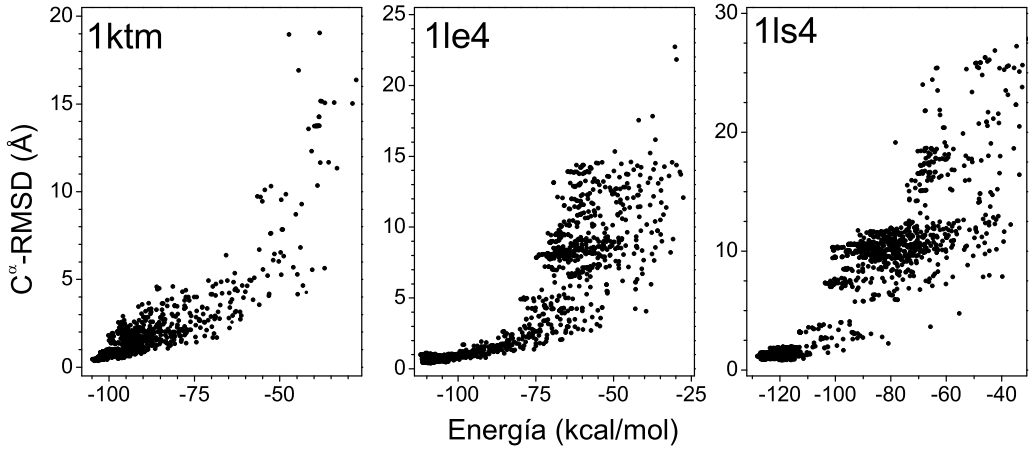


Figura 4.14: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1ktm, 1le4 y 1ls4, en la minimización con el potencial DFIRE-SCM.

energética. Estos resultados son congruentes con la evaluación realizada por miembros del laboratorio de Zhou[136], basada asímismo en minimizaciones.

Otro aspecto satisfactorio de los resultados que obtenemos con este potencial es la definición de la superficie de energía. En las Figuras 4.13 y 4.14 mostramos la representación de energía frente a $RMSD$ de las estructuras de menor energía en algún estadio de la optimización. En todos los casos, el muestreo converge suavemente hacia un mínimo

energético global único. En ninguna de las representaciones aparece un desdoblamiento del mínimo que pueda hacer que la búsqueda se quede estancada en mínimos locales, al contrario de lo que hemos observado con otros potenciales.

4.5. Resumen del Capítulo y conclusiones

En este Capítulo hemos estudiado una serie de potenciales basados en estructuras para la interacción hidrófoba de proteínas. Esta interacción es considerada, generalmente, la fuerza que dirige el plegamiento de proteínas[24]. Para tratar de reproducir sus efectos, muchos grupos de investigación han desarrollado potenciales utilizando diversas aproximaciones. Quizás la más popular entre todas ellas para la comunidad simuladora sea la de elaborar potenciales basados en estructuras, utilizando la gran cantidad de información disponible en el Protein Data Bank[56,57]. Para este estudio hemos seleccionado tres potenciales representativos de distintas aproximaciones que permiten obtener términos de energía entre pares de residuos a partir de estructuras de proteínas. El primero de los potenciales, y el más sencillo de todos ellos, es el potencial de Nanias[100]. Es un potencial basado en la matriz de contactos de Miyazawa y Jernigan[88], obtenida utilizando la aproximación cuasiquímica de Bethe[122]. A este potencial, Nanias *et al.* le añaden una dependencia con la distancia de tipo Lennard-Jones para poder utilizarlo en simulación. El segundo de los potenciales, el TE-13 de Tobi y Elber[93], es un potencial obtenido a partir de la resolución de inecuaciones. Como hemos explicado, este potencial tiene una dependencia con la distancia dividida en 13 intervalos y una forma funcional muy libre, derivada de la aproximación matemática empleada por los autores. El tercero de los potenciales es el potencial DFIRE-SCM, de Zhou *et al.*[123]. Se trata de un potencial estadístico, como el de Miyazawa y Jernigan, aunque en este caso el estado de referencia es el denominado "gas ideal finito"[96,137]. Por otra parte, el potencial DFIRE-SCM como

el TE-13, tiene una dependencia con la distancia dividida en una serie de intervalos, 20 en este caso. Bien por su sencillez, en el caso del potencial de Nanias, o por su éxito en otros estudios comparativos[93,123,136], hemos considerado interesante utilizar estos tres potenciales en nuestra evaluación.

Para poner a prueba la capacidad de los potenciales de generar una superficie de energía potencial que tenga el mínimo en la conformación nativa, hemos utilizado la estrategia evolutiva. Como hemos visto en los Capítulos 2 y 3, nuestro método es capaz de encontrar el mínimo energético en superficies de energía bien definidas. En este caso hemos realizado pequeños cambios en la codificación del algoritmo. Esto supone una reducción de los grados de libertad sobre los que se realiza la búsqueda, por lo que hemos tenido que reevaluar la capacidad del algoritmo de muestrear distintas conformaciones para una determinada proteína. Para ello, hemos utilizado una función de potencial uniformemente atractiva. En concreto hemos utilizado el radio de giro, cuya minimización dirige hacia el colapso de los fragmentos que consideramos de la proteína. Para proteínas de distinta complejidad hemos comprobado esa capacidad de alcanzar distintas conformaciones de la proteína. Esto permite, al hacer un estudio con un potencial de interacción realista, atribuir el éxito o el fracaso al componente de secuencia, y no a la contribución de colapso inespecífico de la interacción hidrófoba.

A continuación, hemos realizado experimentos de minimización para los tres potenciales descritos. Para ello hemos seleccionado proteínas de tipo todo α. Como dijimos en el Capítulo 3, las hélices α que forman estas proteínas se mantienen unidas entre sí debido, fundamentalmente, a interacciones hidrófobas. Con nuestro método, en el que se explora el espacio conformacional sobre disposiciones geométricas de fragmentos rígidos, podemos buscar el empaquetamiento de menor energía de las hélices, y comprobar si corresponde a la conformación nativa.

Los resultados obtenidos con el potencial de Nanias son considerablemente peores

que con los otros dos potenciales. Hemos comprobado que esto se debe en gran medida a la parte repulsiva de la función energética propuesta por los autores. Debido a su hábito, que alcanza rápidamente valores de energía positivos muy grandes para distancias inferiores a 6 Å, para muchas proteínas la energía de la conformación nativa es muy desfavorable. También el valor de σ es problemático, por obviar cualquier consideración sobre el tamaño o el tipo de residuo. Aun así, el potencial mantiene algunas de las buenas propiedades de la matriz de Miyazawa y Jernigan, como hemos visto para proteínas de tres fragmentos en el modelo.

El potencial TE-13 funciona mucho mejor, debido a la cuidada parametrización del potencial. Sin embargo, para que el potencial tenga este buen comportamiento hemos tenido que introducir correcciones para hacerlo compatible con simulaciones del plegamiento. Se requiere una parte repulsiva entre carbonos-α y entre centroides de cadenas laterales para evitar conformaciones en las que los fragmentos de la proteína solapen. Esta parte repulsiva no era necesaria en aproximaciones basadas en quimeras por utilizarse únicamente conformaciones estáticas de la proteína. Sin embargo, son imprescindibles en minimizaciones o simulaciones del plegamiento. Con este potencial, a menudo los mínimos energéticos tienen valores de $RMSD$ próximos a 1 Å, lo cual implica una gran semejanza con la estructura nativa. Aun así, hemos observado una tendencia a que la definición de los mínimos corresponda a un grupo más o menos disperso de conformaciones. Esta definición irregular de la superficie de energía favorece que las minimizaciones queden atrapadas en mínimos locales, como sucede en algunos casos. Pensamos que esto se debe a la forma funcional del potencial que ya hemos comentado. En la representación de la energía entre un par de aminoácidos frente a la distancia se producen saltos de hasta un orden de magnitud entre un intervalo y otro vecino. Al tratar con conformaciones de una proteína, muchos residuos contribuyen a la energía global. Debido a los grandes saltos en la energía puede ser cuestión de azar que se encuentre o no el mínimo

global. Aunque el potencial tiene tendencia a definir correctamente el núcleo hidrófobo de la proteína, hemos visto algunos casos (1i6z, 1ls4) en los que le resulta difícil diferenciar entre empaquetamientos alternativos de los fragmentos rígidos. Así, hemos obtenido mínimos energéticos semejantes a las conformaciones de la proteína que obteníamos con un potencial inespecífico.

Finalmente, el que mejor comportamiento tiene entre los tres potenciales es el DFIRE-SCM. También en este caso hemos tenido que añadir una contribución repulsiva entre carbonos-α y centros de cadenas laterales, igual que con el potencial TE-13. Una vez hecho esto, para todas las proteínas estudiadas hemos encontrado el mínimo en la conformación nativa de la proteína. El estado de referencia no interaccionante definido por los autores parece ser el más apropiado para capturar las tendencias específicas de los distintos pares de residuos. Con este potencial, para los distintos pares de residuos, los saltos energéticos de un intervalo al siguiente en la escala de distancias son mucho más suaves que con el TE-13. Esto hace que la solución sea mucho más fácilmente alcanzable en nuestro estudio de minimización. En todos los casos hemos observado que la superficie de energía explorada dirige inequívocamente hacia el mínimo energético.

Con esta parte de la Tesis contribuimos a la necesidad de validar potenciales de plegamiento obtenidos con distintas aproximaciones[51,55]. Entre los potenciales evaluados, el DFIRE-SCM del laboratorio de Zhou ofrece resultados superiores a los otros dos. Por este motivo, será el que utilicemos cuando llevemos a cabo la evaluación conjunta de varias contribuciones a la energía global de la proteína.

Capítulo 5

Modelos para enlaces de hidrógeno en el esqueleto de proteínas

En este capítulo estudiamos una serie de modelos que tratan de reproducir el efecto de los enlaces de hidrógeno en el esqueleto de proteínas. Un enlace de hidrógeno se forma cuando dos átomos electronegativos D y A compiten por un mismo átomo de hidrógeno[1]:

$$\text{—D—H} \cdots \text{A—}$$

El átomo de hidrógeno está formalmente unido por enlace covalente al átomo donador D, pero interacciona también con el otro, el átomo aceptor A. Podemos describir de forma sencilla el enlace de hidrógeno como una interacción dipolo-dipolo, en la que están implicados el dipolo formado por el átomo donador D y el átomo de hidrógeno, y el dipolo formado por el átomo aceptor A y su base[138]. En proteínas, los enlaces de hidrógeno que aparecen con más frecuencia (más de dos terceras partes del total) son los que se forman entre grupos carbonilo y amino del esqueleto. Estos enlaces de hidrógeno estabilizan los elementos de estructura secundaria, hélices α y láminas β, en los que están implicados la mayoría de residuos de la proteína[26]. La longitud de este tipo de enlace de hidrógeno oscila entre 1.9 y 2 Å y su fortaleza, entre 3 y 10 kcal/mol (entre

10 y 40 kJ/mol, aproximadamente).

La importancia del enlace de hidrógeno en el plegamiento y la estabilidad de proteínas ha sido largamente discutida. Mirsky y Pauling sugirieron en los años 30 del pasado siglo que podían constituir la fuerza dominante en el plegamiento[139]. Estos estudios culminaron con el descubrimiento, por Pauling y Corey, de los principales tipos de estructura secundaria, que como hemos mencionado están estabilizados por enlaces de hidrógeno entre átomos del esqueleto de la proteína[140–143]. Hoy no se mantiene la idea de que los enlaces de hidrógeno sean la fuerza que dirige el plegamiento[24], pero siguen siendo considerados muy importantes por las restricciones energéticas y geométricas que imponen a la estructura tridimensional. Un debate especialmente controvertido es el de en qué medida suponen una estabilización o una desestabilización con respecto al estado desplegado de una proteína[25,27,28].

Para calcular la contribución a la estabilidad de los enlaces de hidrógeno en el estudio teórico de macromoléculas biológicas se han utilizado distintas aproximaciones[138]. Una opción es emplear campos de fuerza de mecánica molecular. En estos campos de fuerza normalmente se usa una combinación de términos coulombianos y de Lennard-Jones, similar al que describe todas las interacciones no enlazantes. Como hemos comentado, los campos de fuerza permiten plegar pequeñas proteínas mediante simulaciones de dinámica molecular consumiendo una gran cantidad de recursos computacionales[144]. Una alternativa es construir potenciales empíricos a partir de la información experimental disponible. Estos modelos tratan de capturar las características promedio de la geometría de los enlaces de hidrógeno en estructuras de proteínas. Los potenciales empíricos pueden ser utilizados junto a términos físicos que reflejen la interacción electrostática, de van der Waals y la solvatación. En este segundo tipo de aproximación se encuadrarían los modelos minimalistas para el enlace de hidrógeno. Los modelos minimalistas usan representaciones de grano grueso de la proteína. Por su simplicidad resultan

muy apropiados para la simulación de procesos como el plegamiento que requieren un gran esfuerzo computacional. En este estudio nos centramos en tres de estos modelos desarrollados por distintos grupos de investigación. A partir de esta comparativa podemos determinar cómo afectan las simplificaciones de cada modelo a la manera en que reproducen los efectos de los enlaces de hidrógeno.

5.1. Modelos de interacción: Irbäck, Chen y Kolinski

A continuación se describen los modelos seleccionados para el estudio de la contribución del enlace de hidrógeno. Los tres modelos han sido tomados de estudios en los que se consideraban otras contribuciones a la energía para la simulación del plegamiento. En este estudio evaluamos independientemente la contribución del modelo de enlace de hidrógeno. Una definición más completa de cada uno de los modelos puede encontrarse en los artículos citados.

5.1.1. Modelo de Irbäck

El primero de los modelos para enlaces de hidrógeno que consideramos es el modelo de Irbäck. Anders Irbäck *et al.* lo han venido utilizando en una serie de trabajos con objeto de obtener información sobre la termodinámica y la dinámica del proceso de plegamiento para proteínas pequeñas[102,145–155]. En la versión original del modelo, cada residuo de la proteína está representado por cinco o seis átomos[102]: los cinco átomos del esqueleto —C^{α}, C', O, N y H—, y para todos los aminoácidos excepto Gly, el átomo de la cadena lateral que está unido al carbono-α, el C^{β}. En los estudios citados el método utilizado es el templado simulado, y los grados de libertad de la cadena polipeptídica sobre los que se realiza el muestreo conformacional son los ángulos de Ramachandran, ϕ y ψ.

La función de energía utilizada por los autores consta de cuatro contribuciones,

entre las que se encuentra la de los enlaces de hidrógeno del esqueleto de la proteína. El término correspondiente a este tipo de interacción viene expresado por

$$E_{hb} = \varepsilon_{hb} \sum_{ij} u(r_{ij}) \nu(\alpha_{ij}, \beta_{ij}). \tag{5.1}$$

Según esta ecuación, la energía asociada a los enlaces de hidrógeno, que recibe un peso ε_{hb} sobre la energía total, se calcula como suma sobre todos los pares ij de residuos de un producto de dos términos, $u(r_{ij})$ y $\nu(\alpha_{ij}, \beta_{ij})$. El primero de los términos, $u(r_{ij})$, es dependiente de la distancia r_{ij} entre el átomo de hidrógeno del residuo i y el átomo de oxígeno del residuo j, y tiene una forma funcional de tipo Lennard-Jones,

$$u(r_{ij}) = 5 \left(\frac{\sigma_{hb}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{hb}}{r_{ij}} \right)^{10}, \tag{5.2}$$

donde σ_{hb} es igual a 2 Å. Este valor para la distancia de equilibrio en el potencial se corresponde con el valor promedio de la distancia H-O en enlaces de hidrógeno formados en el esqueleto de proteínas[1]. La distancia de corte para el cálculo de la energía es r_{ij}=4.5 Å.

El segundo término, $\nu(\alpha_{ij}, \beta_{ij})$, expresa la dependencia angular de la energía:

$$\nu(\alpha_{ij}, \beta_{ij}) = \begin{cases} \cos^2\alpha_{ij} \cos^2\beta_{ij}, & \text{cuando } \alpha_{ij}, \beta_{ij} > 90^\circ \\ 0, & \text{en los casos restantes} \end{cases} \tag{5.3}$$

En la Figura 5.1 mostramos una representación de los ángulos que encontramos en esta ecuación: α_{ij}, entre los átomos N_i, H_i y O_j, y β_{ij}, entre los átomos H_i, O_j y C'_j. Las funciones trigonométricas que se emplean para definir la dependencia con α_{ij} y β_{ij} son de tipo coseno. Por tanto, la situación de mínima energía en el modelo de Irbäck para un par ij de residuos se encuentra cuando la distancia H_i-O_j es de 2 Å y los átomos C'

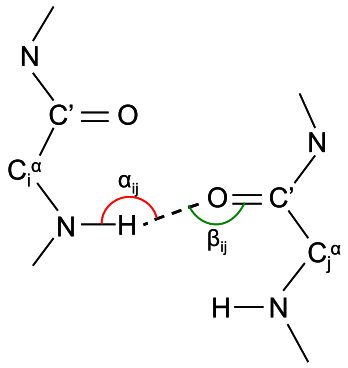


Figura 5.1: Enlace de hidrógeno entre residuos del esqueleto de dos fragmentos de una proteína. Mostramos los ángulos α_{ij} y β_{ij} que se utilizan en el modelo de Irbäck para calcular el término dependiente de la orientación.

y O del residuo j y H y N del residuo i están alineados.

En nuestra implementación del modelo, sólo consideramos la contribución que hemos descrito. Al ser esta la única contribución a la estabilidad que tenemos en cuenta, prescindimos del factor de peso ε_{hb} de la Ecuación 5.1. Para el cálculo de la energía de una determinada conformación, utilizamos las coordenadas atómicas de los cinco átomos del esqueleto de cada aminoácido, sin considerar el efecto de las cadenas laterales. Para evitar que el muestreo conformacional pase por estructuras en la que los átomos estén interpenetrados, introducimos un término de volumen excluido entre carbonos-α, a una distancia de 3.8 Å. Como en nuestro muestreo muchos grados de libertad están congelados, no es necesario que cada átomo del modelo tenga una repulsión estérica individual como en los trabajos de Irbäck *et al.*

Consideramos el modelo de Irbäck como referencia con la que comparar modelos más simplificados para el enlace de hidrógeno del esqueleto de proteínas. El tipo de funcionalidad que utiliza para esta contribución viene avalado por su semejanza con las ecuaciones que se utilizaban para el enlace de hidrógeno en algunos campos de fuerza de mecánica molecular[138]. Por ejemplo, en la versión de 1984 de AMBER[156] se incluía un término específico para el enlace de hidrógeno, de tipo Lennard-Jones, dependiente

de la distancia r_{ij} entre H y O:

$$u(r_{ij}) = \left(\frac{C_{ij}}{{r_{ij}}^{12}}\right) - \left(\frac{D_{ij}}{{r_{ij}}^{10}}\right), \tag{5.4}$$

con parámetros empiricos ajustables C_{ij} and D_{ij}. Por otra parte, en la versión original de CHARMM[157] el término de enlace de hidrógeno también era dependiente de la distancia, con forma funcional de tipo Lennard-Jones. Además, este término estaba modulado por una función angular semejante a la del modelo Irbäck: $\cos^m\alpha_{ij}\cos^n\beta_{ij}$. El parecido de las ecuaciones Irbäck *et al.* con estos dos campos de fuerza respalda nuestra elección del modelo de Irbäck como referencia.

5.1.2. Modelo de Chen

Mucho más simplificado que el modelo de Irbäck es el desarrollado por Jeff Z.Y. Chen e Hideo Imamura. Chen e Imamura parten de un modelo de polímero muy simplificado que les permite reproducir las estructuras que forma el esqueleto de proteínas[104,158–161]. Utiliza sólo tres centros de interacción por aminoácido, cuyas estructuras de mínima energía corresponden a hélices y láminas como las de la Figura 5.2[158]. Realizando pequeños ajustes, el modelo les ha permitido estudiar la interconversión entre los dos tipos de estructura secundaria[160]. En una versión más reciente, los autores han ampliado el modelo añadiendo una contribución de tipo hidrófobo al potencial, para estudiar la estabilidad y la dinámica en el plegamiento de horquillas β[104,161]. En todos sus trabajos han utilizado el método de Monte Carlo para la simulación del templado a diferentes temperaturas.

El punto de partida del modelo es una cadena homopolimérica, en la que las posiciones de los monómeros se corresponden con las de los carbonos-α de una proteína[158]. Las coordenadas de los monómeros se utilizan para calcular las posiciones de dos cen-

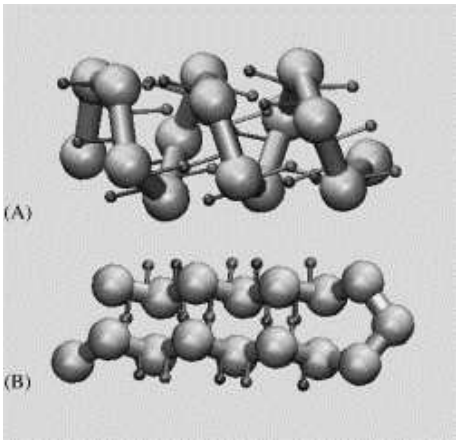


Figura 5.2: Estructuras de tipo helice α (A) y lámina β (B) que aparecen como mínimos energéticos en el modelo de Chen. Cada residuo viene representado por su carbono-α —esferas grandes— y dos átomos virtuales para la interacción de enlace de hidrógeno—esferas pequeñas.

tros de interacción virtuales, que se requieren para definir los enlaces de hidrógeno. Si consideramos el monómero i, cuya posición viene representada por el vector $\mathbf{r_i}$, se puede obtener la posición de su oxígeno virtual (O$'$) mediante

$$\mathbf{r_i}^{(\mathrm{O}')} = \frac{1}{3}\left(\mathbf{r_{i+1}} - \mathbf{r_i}\right) + 0{,}6l\mathbf{n_i}. \tag{5.5}$$

Análogamente, la posición del hidrógeno virtual (H$'$) correspondiente al siguiente monómero se calcula como:

$$\mathbf{r_{i+1}}^{(\mathrm{H}')} = \frac{2}{3}\left(\mathbf{r_{i+1}} - \mathbf{r_i}\right) - 0{,}6l\mathbf{n_i}. \tag{5.6}$$

En estas ecuaciones, l es la distancia entre los carbonos-α i e $(i+1)$, y $\mathbf{n_i}$ es un vector unidad normal al plano formado por los monómeros i, $(i+1)$ e $(i+2)$. De esta manera se pueden reconstruir las posiciones de todos los centros de interacción del modelo, exceptuando al primer monómero de la cadena, del que no se puede calcular la posición de H$'$, el penúltimo, del que no se puede reconstruir O$'$, y el último, para el cual no

podemos obtener las posiciones de H′ ni de O′.

En la versión inicial del modelo[158] —en ausencia de contribución hidrófoba—, la energía de una conformación de una proteína se calcula como suma sobre pares de residuos de la energía de interacción entre los centros H'_i y O'_j:

$$E_{hb} = \sum_{ij} u(r_{ij}) \tag{5.7}$$

El sumatorio se extiende a todos los pares ij de monómeros de la cadena, excepto cuando los autores definen un residuo bisagra, que no puede interaccionar. Cuando todos los residuos de la cadena pueden formar enlaces de hidrógeno, el mínimo energético del modelo se localiza en estructuras de tipo helicoidal[158]. Al introducir residuos bisagra se pueden estabilizar también estructuras de tipo lámina β.

La forma funcional para el término $u(r_{ij})$ de la Ecuación (5.7) es, como en el caso de Irbäck, de tipo Lennard-Jones, pero con un desplazamiento que elimina la parte repulsiva del potencial:

$$u(r_{ij}) = 4\varepsilon \left[\left(\frac{\sigma}{r_{ij} + r_0} \right)^{12} - \left(\frac{\sigma}{r_{ij} + r_0} \right)^{6} \right], \tag{5.8}$$

En esta ecuación, r_{ij} es la distancia entre H'_i y O'_j, y σ es el diámetro de volumen excluido entre monómeros, que se calcula como $\sigma = 1.2 \times l$. Asignamos a l el valor del promedio de la distancia entre carbonos-α vecinos en proteínas reales, 3.8 Å, al ser estos los monómeros del modelo. Por tanto, el valor de σ es 4.56 Å. El desplazamiento de la función de Lennard-Jones r_0 se define como $r_0 = 2^{1/6} \times \sigma$. Los autores no definen una distancia de corte para estas interacciones, aunque nosotros la situamos en $(2.5\sigma - r_0)$ para ahorrar tiempo de cálculo en interacciones que contribuyen de manera insignificante a la energía global.

Además de la contribución de enlaces de hidrógeno, tenemos en cuenta el volumen excluido entre carbonos-α, tal y como lo definen los autores[158]. Así, introducimos un potencial de esferas blandas de diámetro σ.

5.1.3. Modelo de Kolinski

Las primeras menciones al modelo CABS de Kolinski *et al.* las encontramos en el trabajo en que se presentaba TOUCHSTONE II, un método integral para la predicción *ab initio* de la estructura de proteínas[162]. Con este método, los autores pretendían representar la proteína de una manera tan fiel a la estructura como fuese posible, pero que a la vez resultase tratable computacionalmente. Además, querían elaborar un campo de fuerza que tuviese el mínimo global en la vecindad de la conformación nativa. Con el modelo CABS intentaban dar respuesta a estos dos requerimientos, como habían hecho con el modelo SICHO en la anterior versión de TOUCHSTONE[109,163].

En la Figura 5.3 mostramos la representación de un fragmento de una proteína según el modelo CABS. Cada residuo viene representado por tres centros: el carbono-α, el carbono-β (para todos los aminoácidos excepto Gly) y el centro de masas de la cadena lateral (excepto para Gly y Ala)[50,103,162,164–167]. Además, en el centro de cada enlace virtual C^{α}-C^{α} se introduce un nuevo centro de interacción, un átomo virtual al que llamaremos PB, como se ve en la Figura. Los carbonos-α de la proteína se situan en los nudos de una red cúbica de alta resolución, mientras que los carbonos-β, los centros de masas de las cadenas laterales y los átomos PB quedan fuera de la red.

La energía de una conformación de la proteína se calcula en el modelo como suma de términos, con contribuciones de corto y largo alcance. Dada la ausencia de detalle atómico a nivel de esqueleto en el modelo, los enlaces de hidrógeno se definen como interacciones direccionales entre carbonos-α. Para una determinada conformación de la proteína, la energía debida a este tipo de interacción se calcula como un sumatorio entre

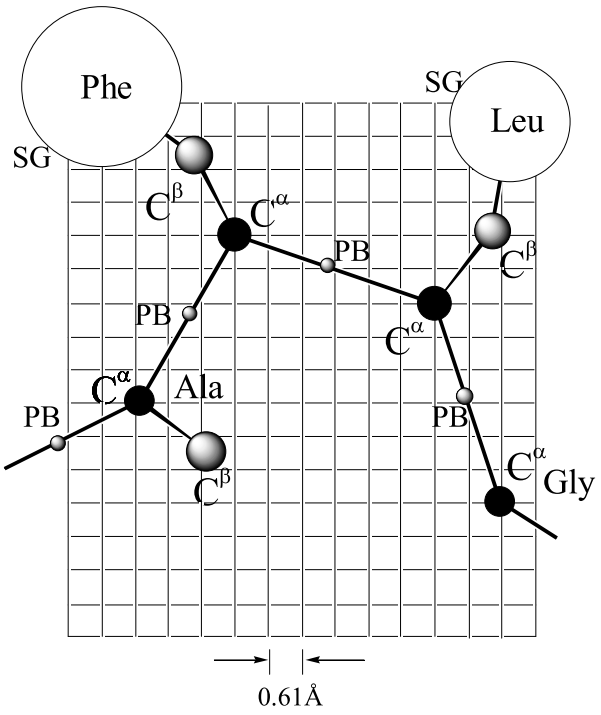


Figura 5.3: Fragmento de una proteína formado por Ala-Phe-Leu-Gly según el modelo de Kolinski. Cada residuo está representado por su C^α, que ocupa un nudo de una red cúbica, su C^β y el centro de masas de la cadena lacteral, SG, ambos fuera de red. Entre cada par de C^αs se representa un átomo virtual PB, también fuera de red.

pares de residuos:

$$E_{hb} = \sum_{ij} u_{ij} \tag{5.9}$$

El término u_{ij} en el modelo corresponde, según los autores, a los enlaces de hidrógeno reales que se establecen entre los residuos $(i, j+1)$. De tal modo que en el modelo de Kolinski hay una renumeración de las interacciones, que permite comparar las que aparecen en el modelo con las de la proteína. El sumatorio de la Ecuación (5.9) se extiende sobre todos los pares ij exceptuando el término $(i, i+4)$. Dada la correlación entre enlaces de hidrógeno reales e interacciones en el modelo, al suprimir los enlaces de hidrógeno $(i, i+4)$ se estaría evitando la estabilización de estructuras con interacciones reales $(i, i+5)$. Así, se impide que se formen estructuras con hélices gruesas, en favor de las interacciones $(i, i+4)$ reales, que estabilizan las hélices α.

Antes de definir el término u_{ij} de la Ecuación (5.9) es necesario introducir una serie de vectores del modelo. Si se considera el residuo i, el vector de enlace $\mathbf{v_i}$ se define como aquel que une los carbonos-α i e $(i+1)$, tal y como mostramos en la Figura 5.4. Una vez definidos los vectores $\mathbf{v_i}$ y $\mathbf{v_{i-1}}$ podemos obtener el vector normalizado bisector del ángulo que forman, $\mathbf{b_i}$, como se indica en la Figura 5.5:

$$\mathbf{b_i} = \frac{\mathbf{v_{i-1}} - \mathbf{v_i}}{\left|\mathbf{v_{i-1}} - \mathbf{v_i}\right|} \tag{5.10}$$

Por último, podemos calcular el vector de enlace de hidrógeno $\mathbf{h_i}$. Este vector tiene módulo 4.6 Å y es ortogonal al plano formado por los vectores de enlace $\mathbf{v_{i-1}}$ y $\mathbf{v_i}$.

Una vez calculados estos vectores y previo al cálculo de la energía, debe comprobarse si se cumplen una serie de restricciones para la formación de un enlace de hidrógeno entre los residuos i y j:

1. Distancia máxima entre carbonos-α de los residuos i y j:

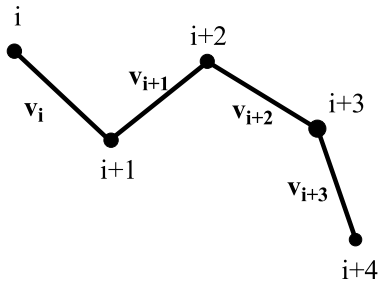


Figura 5.4: Fragmento de una proteína con la definición de los vectores **v** del modelo de Kolinski.

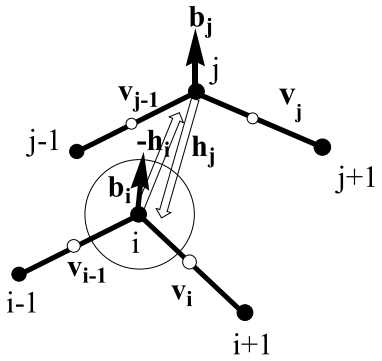


Figura 5.5: Dos fragmentos de una proteína con la definición de los vectores **b** y **h** en el modelo de Kolinski.

$$\left|\mathbf{r_i} - \mathbf{r_j}\right| < 6.1 \text{ Å}$$

2. Restricción direccional en virtud del cual los vectores de enlace de hidrógeno $\mathbf{h_i}$ y $\mathbf{h_j}$ no deben desviarse de la orientación paralela o antiparalela más de 40°:

$$\left|\mathbf{h_i} \cdot \mathbf{h_j}\right| > 16 \text{ Å}^2$$

3. Condición para que la orientación relativa de los dos fragmentos del esqueleto de la proteína que vayan a formar enlace de hidrógeno sea aproximadadamente paralela o antiparalela:

$$[(\mathbf{v_{i-1}} \cdot \mathbf{v_j}) < 0] \text{ y } [(\mathbf{v_{j-1}} \cdot \mathbf{v_i}) < 0], \text{ o bien: } [(\mathbf{v_i} \cdot \mathbf{v_j}) > 0] \text{ y } [(\mathbf{v_{j-1}} \cdot \mathbf{v_{i-1}}) > 0]$$

4. Semejanza entre el vector $\mathbf{r_{ij}}$ y los vectores de enlace $\mathbf{h_i}$ y $\mathbf{h_j}$:

$$\left|\,(\mathbf{r_i} - \mathbf{r_j}) - \mathbf{h_i}\right| < 1.83 \text{ Å, o bien: } \left|\,(\mathbf{r_i} - \mathbf{r_j}) - \mathbf{h_j}\right| < 1.83 \text{ Å}$$

Una vez comprobadas las restricciones, la energía de la interacción se calcula como suma de términos:

$$u_{ij} = \delta_{ij}^h \varepsilon^h u_{ij}^h + \delta_{ij}^\gamma \varepsilon^\gamma u_{ij}^\gamma \tag{5.11}$$

El primero de ellos, u_{ij}^h, contribuye a la energía global ($\delta_{ij}^h \neq 0$) cuando se cumplen las cuatro restricciones. La cuarta de las restricciones se formula para los dos vectores de enlace de hidrógeno, es decir $\mathbf{h_i}$ y $\mathbf{h_j}$. Cuando esta restricción se cumple para uno de los dos vectores, el factor δ_{ij}^h vale 1, y cuando se cumple para los dos vectores, δ_{ij}^h vale 2. Cuando no se cumplen las cuatro restricciones $\delta_{ij}^h = 0$. El término u_{ij}^h se calcula como

$$u_{ij}^h = 0.5 + \left(\frac{4.25}{\text{máx}\left[4.25, \text{mín}\left(6.01, r_{pp}\right)\right]}\right)^4 + \left(\frac{4.25}{\text{máx}\left[4.25, \text{mín}\left(6.01, r_{qq}\right)\right]}\right)^4, \qquad (5.12)$$

donde r_{pp} y r_{qq} son las distancias entre átomos PB enfrentados. En la Ecuación (5.11), u_{ij}^h está multiplicado por ε^h, que vale -1.25. El producto $\varepsilon^h u_{ij}^h$ alcanza su valor mínimo para distancias r_{pp} y r_{qq} pequeñas. Esto supone que este término de la energía es más favorable para aquellas conformaciones en las que los fragmentos de cadena en los que se encuentran los residuos i y j están próximos entre sí.

El factor δ_{ij}^γ por el que está multiplicado el segundo término vale 1 cuando para el par ij se cumplen las tres primeras restricciones. Si no es así, δ_{ij}^γ es 0. El valor que recibe u_{ij}^γ viene dado por

$$\begin{aligned} u_{ij}^\gamma = 2 - \text{máx}\left[\left(\mathbf{b_i}\cdot\left(\mathbf{r_i}-\mathbf{r_j}\right)/6.1\right)^2, 0.125\right] - \\ - \text{máx}\left[\left(\mathbf{b_j}\cdot\left(\mathbf{r_i}-\mathbf{r_j}\right)/6.1\right)^2, 0.125\right] \end{aligned} \qquad (5.13)$$

El factor de peso asociado a u_{ij}^γ es ε^γ, cuyo valor es -0.25. La energía mínima debida a esta contribución corresponde a conformaciones en las que los términos $\mathbf{b_i}\cdot\left(\mathbf{r_i}-\mathbf{r_j}\right)$ y $\mathbf{b_j}\cdot\left(\mathbf{r_i}-\mathbf{r_j}\right)$ tienen valores bajos. Esto se favorece cuando los carbonos-α de los residuos i y j están próximos entre sí, y cuando el ángulo que forma el vector que une los carbonos-α con cada uno de los vectores bisectores, $\mathbf{b_i}$ y $\mathbf{b_j}$, es próximo a 90°.

Debido a los factores de peso ε^h y ε^γ y los valores mínimos que pueden alcanzar u_{ij}^h y u_{ij}^γ, la contribución de cada uno de los términos de la Ecuación (5.11) a la energía

total es diferente. El valor mínimo que puede adoptar el producto $\delta_{ij}^{h}\varepsilon^{h}u_{ij}^{h}$ es -3.125, mientras que el del producto $\delta_{ij}^{\gamma}\varepsilon^{\gamma}u_{ij}^{\gamma}$ puede valer hasta -0.438.

Además de la componente de los enlaces de hidrógeno, en nuestra implementación del modelo de Kolinski hemos incluido repulsiones de volumen excluido semejantes a las que los autores describen en sus trabajos[103]. Hemos utilizado un potencial de esferas blandas entre carbonos-α genérico, de diámetro $\phi_{C^{\alpha}-C^{\alpha}}$=3.8 Å, y otro entre carbonos-α y átomos virtuales PB, de diámetro $\phi_{C^{\alpha}-PB}$ =4.2 Å.

5.2. Eficiencia en la representación de los enlaces de hidrógeno en 2gb1

Cada uno de los modelos que hemos introducido en el apartado anterior reproduce de una manera particular el efecto de los enlaces de hidrógeno del esqueleto. Para comprender más hondamente cómo cada modelo trata de captar esas interacciones, hemos estudiado con cada uno de ellos la conformación nativa del dominio de unión a inmunoglobulina de la proteína G de estreptococo, cuyo código PDB es 2gb1. Hemos seleccionado este dominio de 56 aminoácidos porque en su estructura nativa encontramos los dos tipos principales de estructura secundaria, hélices α y láminas β. Como hemos comentado repetidas veces, estas estructuras están estabilizadas por enlaces de hidrógeno entre grupos del esqueleto peptídico. Por ello, 2gb1 puede ser un ejemplo de proteína muy apropiado para conocer el comportamiento de los distintos modelos.

Para cada uno de los modelos, hemos tomado del archivo PDB de 2gb1 la información necesaria para el cálculo de la energía. En el caso del potencial de Irbäck, se utilizan las coordenadas de todos los átomos del esqueleto. En la Figura 5.6 (a) se muestra la representación del esqueleto de la proteína que utiliza este modelo. Para los modelos de Chen y Kolinski sólo son necesarias las coordenadas de carbonos-α, que mostramos

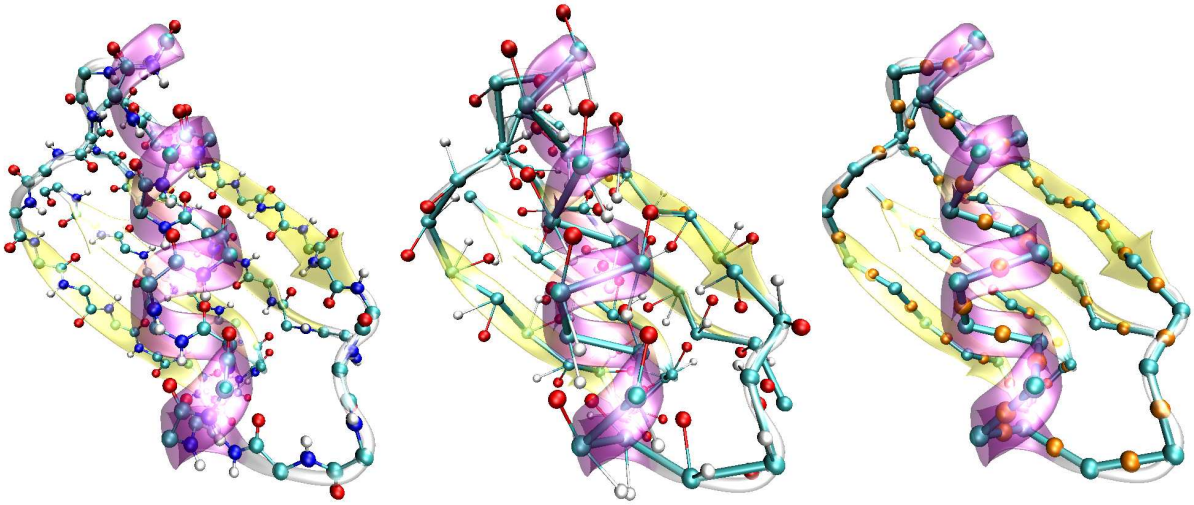


Figura 5.6: El dominio 2gb1 según los tres modelos para la interacción por enlace de hidrógeno: (a) modelo de Irbäck, (b) modelo de Chen, y (c) modelo de Kolinski. Superpuesta a cada una de las estructuras se muestra la representación de las hélices α de la proteína en morado y las láminas β en amarillo.

en la Figura 5.6 (b) y (c). A partir de ellas se obtienen las de los restantes centros de interacción para el enlace de hidrógeno: los átomos virtuales H$'$ y O$'$ en el caso del modelo de Chen y los centros PB en el modelo de Kolinski. Utilizando estas coordenadas hemos calculado los mapas de energía, que mostramos en la Figura 5.7 para los distintos modelos. Estos mapas permiten ver cómo cada uno de ellos representa las interacciones nativas para esta proteína.

5.2.1. Modelo de Irbäck

Como hemos comentado en la Sección 5.1.1, utilizamos el modelo de Irbäck como referencia en este estudio de modelos de enlace de hidrógeno. En la Figura 5.6 (a) mostramos una representación de la conformación nativa de la proteína con el nivel de detalle del modelo de Irbäck para el esqueleto de la proteína. Para cada uno de los residuos representamos todos los átomos del esqueleto, N, H, C$^{\alpha}$, C$'$ y O, cuyas coordenadas están

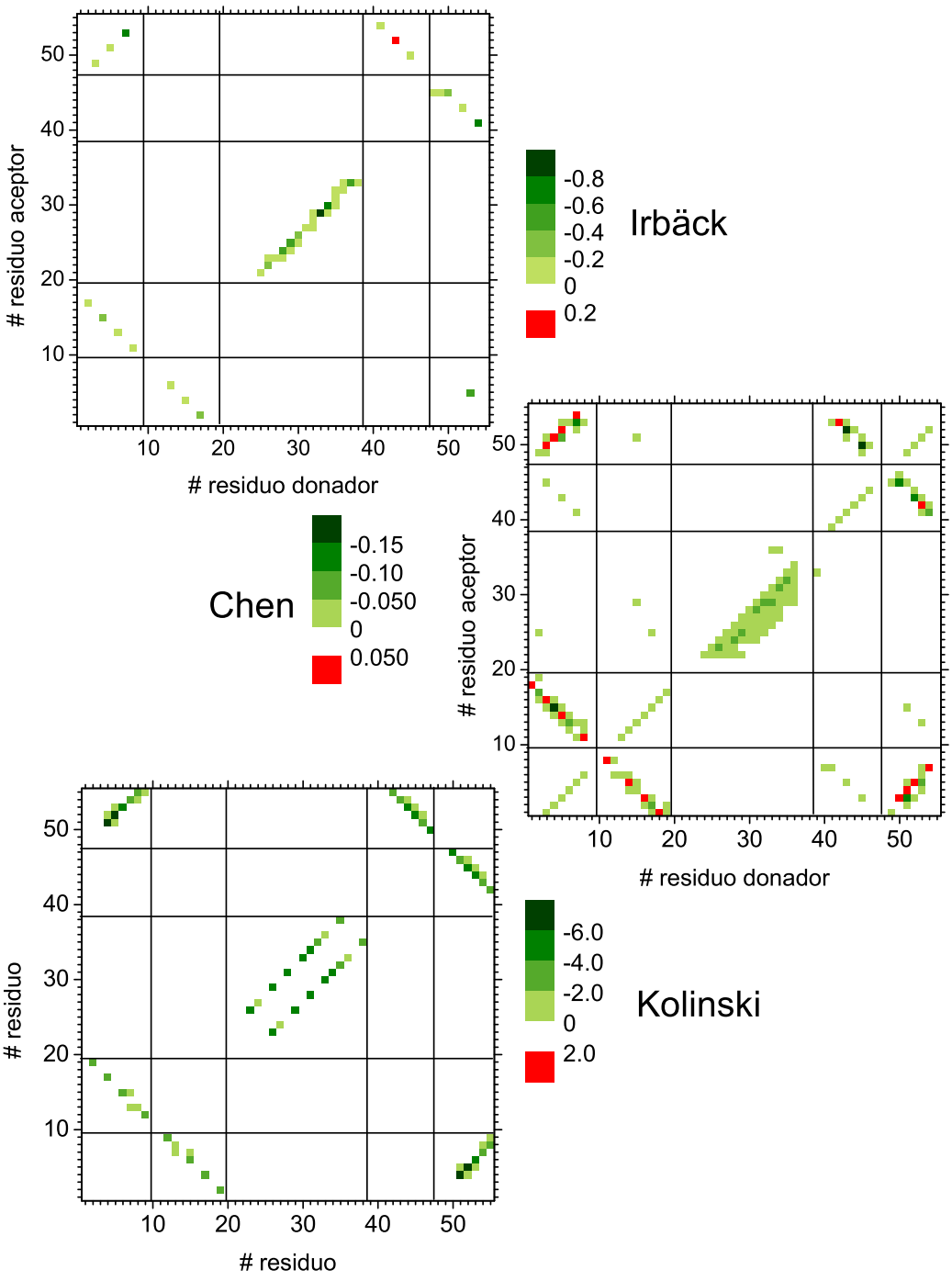


Figura 5.7: Mapas de energía de la conformación nativa de 2gb1 para los tres modelos de interacción de tipo enlace de hidrógeno. Las escalas energéticas son arbitrarias para los tres modelos e independientes entre sí.

tomadas directamente del archivo PDB. No se incluyen los carbonos-β, que también forman parte del modelo de Irbäck, porque aquí sólo vamos a considerar la contribución de enlaces de hidrógeno del esqueleto. En la representación del modelo mostramos superpuestos los diagramas correspondientes a los dominios con distinto tipo de estructura secundaria.

A partir de las coordenadas de los átomos podemos obtener la energía de cada una de las interacciones de enlace de hidrógeno. La energía que el modelo asigna al conjunto de enlaces de hidrógeno de la proteína es -7 (en unidades arbitrarias). Hemos representado las interacciones entre pares de centros en un mapa de energías que mostramos en la Figura 5.7. Dado que el modelo tiene detalle atómico, puede distinguirse entre aquellos casos en que un residuo actúa como donador de aquellos en que actúa como aceptor. En los ejes de la Figura se muestran los índices correspondientes a la posición de cada uno de los residuos en la secuencia de la proteína; en el eje de abscisas, cuando actúan como donadores, y en el de ordenadas, cuando intervienen como aceptores. Hemos delimitado con líneas negras las regiones con distinto tipo de estructura secundaria en la conformación nativa de la proteína. Empezando por el principio de la secuencia, nos encontramos las hebras $\beta 1$ y $\beta 2$, después la hélice α, y finalmente las hebras $\beta 3$ y $\beta 4$. Cada uno de los puntos coloreados corresponde a una interacción.

Tanto las láminas β como las hélices α se caracterizan por un patrón de enlaces de hidrógeno particular de los distintos tipos de estructura secundaria[25]. Por ejemplo, en las láminas antiparalelas un residuo i interacciona con otro j como donador de hidrógeno y también como aceptor. En el caso de las láminas paralelas, un residuo i interacciona con otro j como donador, pero interacciona como aceptor con el $(j+1)$. El modelo de Irbäck es capaz de reproducir estos patrones de interacción para 2gb1. En la estructura terciaria de esta, las cuatro hebras forman parte de una misma lámina. En esta lámina, las hebras $\beta 1$ y $\beta 2$ son antiparalelas, así como la $\beta 3$ con la $\beta 4$, mientras que $\beta 1$ y $\beta 4$

son paralelas entre sí. Si nos fijamos en los enlaces de hidrógeno de las hebras $\beta 1$ y $\beta 2$ en la Figura 5.7, vemos que un mismo residuo interacciona como donador y como aceptor, mientras el siguiente no forma ningún enlace de hidrógeno ni como donador ni como aceptor. Lo mismo sucede entre hebras $\beta 3$ y $\beta 4$, también antiparalelas. Así, se forman unas "diagonales" de contactos, en las que las interacciones se van alternando. Las hebras $\beta 1$ y $\beta 4$, por el contrario, son paralelas entre sí, y en el caso de 2gb1 aparecen peor representadas en el modelo de Irbäck. Sólo en un caso podemos ver cómo capta el patrón característico de las láminas paralelas: el residuo 53 interacciona como donador con el 5, mientras que es aceptor de hidrógeno del 7. La peor representación del patrón de interacciones de las láminas paralelas se debe al menor tamaño de los fragmentos correspondientes. Aún así, en la región en que los residuos de la hebra $\beta 1$ interaccionan como donadores con los de la hebra $\beta 4$, se puede ver perfectamente la correspondiente "diagonal" de contactos.

Observamos que el modelo también es capaz de identificar situaciones especiales como la del residuo 45. En el mapa vemos que este grupo aceptor está interaccionando simultáneamente con tres grupos donadores: los átomos de hidrógeno de los residuos 48, 49 y 50. La interacción H_{45}-O_{50} corresponde al patrón típico de la lámina antiparalela; las restantes interacciones, no. Este tipo de situaciones en las que un mismo aceptor interacciona a la vez con varios grupos donadores lo encontramos en regiones donde el esqueleto está más doblado que en estructuras regulares. En el caso del O_{45} esta mayor flexibilidad se debe a que los grupos aceptores H_{48} y H_{49} forman parte de un lazo, donde la cadena cambia su sentido de propagación.

El patrón de enlaces de hidrógeno correspondiente a las hélices α es distinto del de las láminas β. En este caso se forman enlaces de hidrógeno entre el grupo carbonilo del residuo i y el amino del residuo $(i+4)$[25]. En la Figura 5.7, la región central del mapa de Irbäck corresponde a las interacciones de la hélice α. Podemos ver que desde el residuo

21 hasta el 33, como aceptor, o desde el 25 hasta el 37, como donador, el modelo localiza el patrón característico de enlaces de hidrógeno. Además, aparecen algunas interacciones fuera de la diagonal definida por (O_i,H_{i+4}), aunque en general son menos intensas que estas.

Finalmente, podemos reseñar que en el mapa obtenido con el modelo de Irbäck para 2gb1 aparece un contacto repulsivo en la interacción H_{43}-O_{52} (en rojo en la Figura). Esta repulsión aparece debido a que la distancia entre centros es 1.78 Å, inferior al punto de corte de la función del potencial con el eje de abscisas, que se encuentra a 1.83 Å. Así, a esta interacción el modelo le asigna una energía positiva, de 0.73 u.a. Este contacto repulsivo pone de manifiesto el problema de cualquier parametrización sencilla. Los autores han escogido un único valor para σ, independiente de los contextos químicos particulares. Al tratarse de un valor promedio hay algunos casos para los que no es válido.

5.2.2. Modelo de Chen

En nuestra implementación del modelo de Chen, para el cálculo de los enlaces de hidrógeno de la proteína 2gb1, utilizamos las coordenadas de los carbonos-α de su archivo PDB. A partir de ellas, se generan las posiciones de los centros de interacción en el modelo, los átomos virtuales H′ y O′, tal y como hemos indicado en la Sección 5.1.2. Así obtenemos la representación del esqueleto para la proteína, que mostramos en la Figura 5.6 (b), con tres centros por cada residuo.

El primer aspecto que debemos estudiar en este modelo es qué tal funciona la reconstrucción de los centros de interacción. Para ver cómo se correlacionan las posiciones reales y virtuales de los mismos podemos estudiar el valor de $RMSD$ de las posiciones de unos respecto a otros. Este valor, para un tipo de átomo A, que puede ser hidrógeno

u oxígeno, se calcula como:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_{\mathrm{A},\mathbf{i}} - \mathbf{r}_{\mathrm{A}',\mathbf{i}})^2}. \tag{5.14}$$

En esta ecuación, $\mathbf{r}_{\mathrm{A},\mathbf{i}}$ y $\mathbf{r}_{\mathrm{A}',\mathbf{i}}$ son las posiciones de los átomos A y A$'$ correspondientes al residuo i, y N es el número de residuos de la proteína. En el caso de la proteína 2gb1, el valor de $RMSD$ para los átomos de oxígeno es 1.44 Å, y para los átomos de hidrógeno, 1.50 Å. Esta diferencia entre las coordenadas de los átomos reales y los virtuales no debe alarmarnos. Como hemos comentado, en la estructura terciaria de proteínas, la distancia promedio del enlace de hidrógeno del esqueleto, entre los átomos de H y O, es de 2 Å[1]. En cambio, en la ecuación para la energía del modelo de Chen, el mínimo se encuentra para una distancia entre centros virtuales r_{ij}=0. Es razonable, por tanto, que entre los centros de interacción virtuales y reales haya un desplazamiento de cerca de 1 Å, tanto para hidrógenos como para oxígenos.

Para cada residuo i de la proteína, hemos estudiado el valor de la desviación cuadrática entre la posición del átomo virtual y el real, $\left[r_{\mathrm{A},i} - r_{\mathrm{A}',i}\right]^2$, para oxígenos e hidrógenos. En la Figura 5.8 mostramos la distribución de estos valores a lo largo de la secuencia. En la parte superior de la Figura incluimos un diagrama de los fragmentos con distinto tipo de estructura secundaria obtenidas con el algoritmo STRIDE[124], coincidente en lo esencial con la asignación DSSP[168]. A lo largo de la secuencia de la proteína, la desviación cuadrática se mantiene en valores bajos para la mayor parte de los residuos. Sin embargo, en algunas posiciones se alcanzan valores mas altos de desviación cuadrática, tanto para oxígenos como para hidrógenos. Como se puede ver en la Figura, estas posiciones corresponden a residuos situados en los lazos entre fragmentos con estructura secundaria bien definida. Precisamente estas posiciones son en las que los autores del modelo introducen un monómero diferente. En los trabajos originales de

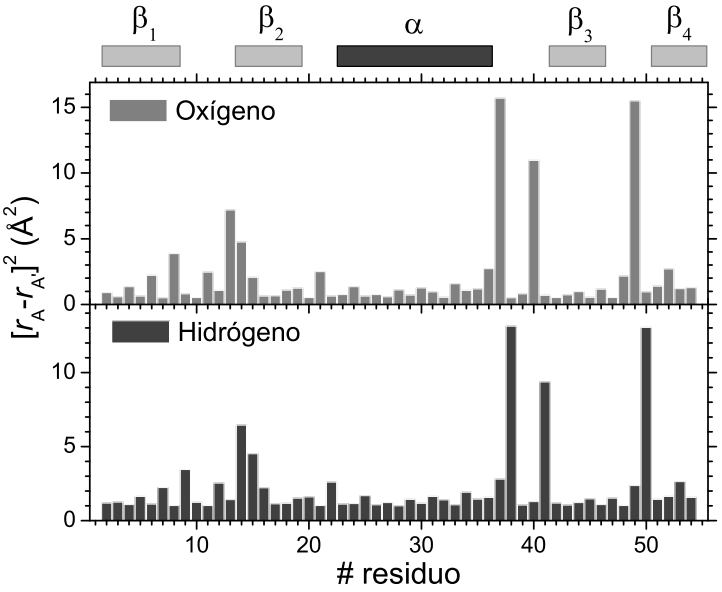


Figura 5.8: Diagrama de columnas para la desviación cuadrática de las posiciones virtuales y reales de átomos de hidrógeno y oxígeno a lo largo de la secuencia de 2gb1. En la parte superior se representan los dominios de estructura secundaria: hélice α y hebras β.

Chen e Imamura[158,160], estos residuos no tenían la capacidad de establecer enlaces de hidrógeno, y hacían las veces de bisagra para que se formasen distintos tipos de estructura secundaria. Por tanto, la mala reconstrucción de estas posiciones de los átomos virtuales no supone un problema para el modelo.

En la Figura 5.7 mostramos el mapa de enlaces de hidrógeno de 2gb1 que hemos obtenido para el modelo de Chen. En este caso, como en el del modelo de Irbäck, cada residuo puede estar interaccionando como aceptor y donador de hidrógeno. Por tanto los ejes del mapa se definen de la misma forma que en el mapa de Irbäck. Sin embargo, entre las dos representaciones se aprecian diferencias importantes. En el mapa obtenido con el modelo de Chen podemos contabilizar un número mucho mayor de interacciones que en el caso de Irbäck. Entre ellas, las de mayor intensidad coinciden con las que ya aparecían en el mapa de Irbäck, es decir, los enlaces de hidrógeno reales o nativos. Pero, además, se forman interacciones más débiles entre centros que en la proteína no

se encuentran formando enlaces de hidrógeno. Este efecto del modelo de Chen se debe a la falta de una dependencia orientacional en la energía. Se puede observar, por ejemplo, que aparecen interacciones dentro de una misma hebra β, en las que se cumple que O$'$ de un residuo i forma un enlace de hidrógeno con H$'$ del residuo $(i+2)$. Al no haber restricciones orientacionales nada impide que se produzcan estas interacciones. Por la misma razón, en la región que corresponde a la hélice α hay muchos más contactos que los $(i, i+4)$ característicos de este tipo de estructura. También ocurre entre hebras de la lámina β entre las que se forman enlaces de hidrógeno nativos. Por ejemplo, entre las hebras $\beta 1$ y $\beta 2$ se forma la denominada diagonal de contactos, los enlaces de hidrógeno alternantes que veíamos con el potencial de Irbäck. Además, siempre con el modelo de Chen, aparecen nuevas interacciones por encima y por debajo de esta diagonal.

Otras interacciones que no se observan en el mapa de enlaces de hidrógeno de Irbäck y sí encontramos en el de Chen son las que se establecen entre la hélice y las hebras $\beta 1$ y $\beta 2$, o entre las hebras $\beta 1$ y $\beta 3$. Estas interacciones aparecen debido a la falta de dependencia angular en la función de la energía, que ya hemos mencionado, y a que no hay una distancia de corte para las interacciones (aunque nosotros asignemos una para evitar el cálculo de contribuciones poco significativas).

Otro aspecto que llama la atención en el mapa obtenido con el potencial de Chen es la gran cantidad de interacciones desfavorables (destacadas en color rojo en la Figura 5.7). En la función de energía de enlace de hidrógeno de Imamura y Chen el mínimo del potencial aparece cuando los centros de interacción solapan. Por tanto, no tiene parte repulsiva. Las repulsiones se deben a la interacción de volumen excluido entre carbonos-α, cuyo diámetro es 4.56 Å. Se producen entre hebras β que interaccionan en la conformación nativa. En general, esos contactos repulsivos se establecen en la misma diagonal que los enlaces de hidrógeno nativos, entre aquellos pares de residuos que no forman este tipo de interacción. Tenemos que pensar, por tanto, que la distancia que

los autores han fijado para el volumen excluido puede ser demasiado grande para ser compatible con estructuras de tipo lámina β reales.

Un último aspecto del modelo que podemos subrayar a partir de la información del mapa es que debido a la reconstrucción incompleta de los centros de interacción virtuales H$'$ y O$'$ hay interacciones que no se pueden calcular. Si nos fijamos en la fila que corresponde a O$'$ del residuo 54 de la proteína, vemos que no establece ninguna interacción aparte de la repulsiva con el residuo 7, de volumen excluido. Esto se debe a que, de acuerdo con el esquema de Chen e Imamura, esa posición no se puede reconstruir. Sin embargo, sí aparecen interacciones para el átomo H$'$ del residuo 54. Esto puede introducir algún sesgo cuando se utilice este potencial en simulaciones reales.

5.2.3. Modelo de Kolinski

En la Figura 5.6 (c) se muestra la representación de la conformación nativa de la proteína 2gb1 tal y como consideramos el modelo de Kolinski. La diferencia principal con la representación que los autores utilizarían es que, en nuestro caso, los carbonos-α ocupan sus posiciones reales en la conformación nativa de la proteína, no las que les corresponderían dentro de una red de alta resolución. Entre cada par de carbonos-α sucesivos en la secuencia se encuentran situados los átomos PB (en naranja en la Figura). Por otra parte, como ya hemos comentado, no tenemos en cuenta los átomos de las cadenas laterales, ya que estamos interesados únicamente en la contribución de enlace de hidrógeno. Así, con este modelo se representan dos centros de interacción por residuo de la proteína.

A partir de esta información, se puede calcular la energía debida a las interacciones de enlace de hidrógeno. Mostramos el mapa resultante de este cálculo en la Figura 5.7. Como en el modelo de Kolinski no se distingue entre el centro donador y el aceptor de cada residuo, los dos ejes significan lo mismo, y el mapa de enlaces de hidrógeno es

simétrico con respecto a la diagonal $x = y$.

En el mapa hay cuatro regiones en las que aparecen interacciones. En este sentido, sucede como en el mapa correspondiente al modelo de Irbäck. Sólo se detectan interacciones entre residuos de la hélice α y entre aquellas hebras de la lámina β que son vecinas en la estructura terciaria. No se encuentran aquí interacciones inespecíficas dentro de un mismo fragmento, ni entre fragmentos que no interaccionan en la conformación nativa, como sucedía con el modelo de Chen.

Por otra parte, el modelo reproduce con más exactitud las interacciones nativas de la hélice que las de la lámina. En la región de la hélice α, las interacciones más importantes son las de la diagonal $(i, i+3)$ (ver Figura 5.7). Una hélice α real está definida por los enlaces de hidrógeno $(i, i+4)$ paralelos a su eje. En el modelo, el patrón se desplaza un residuo debido a la ya mencionada renumeración de las interacciones que los autores sugieren para equiparar su modelo con las interacciones reales[103]. En cuanto a la lámina β, el aspecto más distintivo del mapa es que se pierde el patrón característico de interacciones. De acuerdo con este patrón, en una hebra se alternan los residuos que forman enlaces de hidrógeno con otra, sea la lámina resultante paralela o antiparalela. Al contrario de lo que sucedía con los modelos de Irbäck y Chen, con el de Kolinski no se observa esta pauta. Por ejemplo, todos los residuos de la hebra $\beta 1$ forman interacciones con la hebra $\beta 4$ (ver Figura 5.7). Lo mismo sucede entre la hebra $\beta 3$ y la hebra $\beta 4$. Pensamos que la pérdida del patrón de la lámina no se debe a una incorrección en el modelo. Probablemente, los autores no pretenden reproducir uno a uno los enlaces de hidrógeno, sino contemplar su efecto global.

Otro aspecto importante del mapa es que hay muchas interacciones nativas que el modelo de Kolinski no es capaz de detectar. Esto se debe a las restricciones previas al cálculo de la energía, que pueden ser demasiado estrictas. Podemos constatarlo comparando el mapa de Irbäck —que captura las interacciones "reales"— con el de Kolinski en

la Figura 5.7. Si nos fijamos en la zona de la hélice o en la interacción entre las hebras $\beta 1$ y $\beta 2$, en las correspondientes "diagonales" de enlaces de hidrógeno, vemos que en el mapa de Kolinski aparecen huecos que no encontramos en el de Irbäck. Corresponden a pares de residuos para los cuales no se cumplen una o varias de las restricciones. En los casos en que la energía es más baja en valor absoluto (en color verde claro en la Figura), la restricción que no se cumple es la cuarta, de la que depende el cálculo de la contribución más importante a la energía global. En los que la energía es cero, no se cumplen una o varias de las otras tres restricciones.

Paradójicamente, hay ocasiones en las que residuos no interaccionantes en la proteína sí cumplen las restricciones del modelo de Kolinski. Por ejemplo, vemos en el mapa que el residuo 6 establece dos interacciones, con los residuos 12 y 14, ambas no nativas. Se trata de interacciones de carácter débil, porque ni en el caso (6,12) ni en el (6,14) se cumple la cuarta restricción. Sin embargo, el enlace de hidrógeno nativo (6,13), que sí observamos en el mapa obtenido con el potencial de Irbäck, no se forma debido a que no se verifica $\left|\mathbf{h_6} \cdot \mathbf{h_{13}}\right| > 16$ (su valor para este par de residuos es 13.69). Este es otro ejemplo de la escasa precisión del modelo a la hora de representar interacciones individuales.

5.3. Minimización de la energía de enlace de hidrógeno para proteínas β

Hemos llevado a cabo experimentos de minimización para localizar el mínimo global de energía con los tres modelos de enlace de hidrógeno que hemos descrito. Para ello utilizamos nuestro algoritmo evolutivo. En este caso, la función de mérito del algoritmo, que establece la bondad de las distintas soluciones de una población, es igual a la energía de cada conformación calculada con el modelo correspondiente.

En la Figura 5.9 mostramos el conjunto de proteínas que estudiamos. Todas ellas son péptidos o fragmentos de proteínas con estructura secundaria de tipo lámina β. Hemos escogido este tipo de proteínas para estudiar potenciales de enlace de hidrógeno porque, si dividimos una lámina β en las hebras que la forman, su estabilidad se explica, fundamentalmente, por este tipo de interacciones. Por tanto, dividimos cada una de las proteínas de la Figura en fragmentos de acuerdo con el número de hebras β que tenga en su conformación nativa. Así, el problema de muestreo conformacional queda reducido a una búsqueda sobre los posibles empaquetamientos de las hebras β.

Para cada una de las proteínas, hemos tomado de la base de datos las coordenadas de sus átomos en la conformación nativa. En el caso de las minimizaciones realizadas con el modelo de Irbäck, utilizamos las posiciones de los cinco átomos del esqueleto por cada residuo. Por el contrario, con los modelos de Chen y Kolinski basta con considerar las coordenadas de los carbonos-α y obtener a partir de ellas las posiciones de los centros de interacción virtuales. Para dividir la secuencia en fragmentos, seguimos las indicaciones del PDB y del programa STRIDE [124] de asignación de estructura secundaria. Los residuos que forman parte de los lazos entre hebras β los asignamos a los fragmentos de tal manera que el corte entre fragmentos corresponda al centro de un giro. En todos los casos, la separación entre hebras sucesivas es sólo de un enlace virtual. Por ello, como ya se hizo en la evaluación de poteciales hidrófobos, utilizamos la codificación interna del algoritmo evolutivo, más adecuada para fragmentos conectados entre sí por lazos cortos (ver Capítulo 2).

Hemos implementado el método de minimización para los tres modelos en una serie de programas que se distinguen únicamente en la representación de la proteína y el potencial de interacción. A continuación, hemos ejecutado los correspondientes programas para todas las proteínas. Para los tres modelos y con todas las proteínas, hemos utilizado unos mismos valores de los parámetros del algoritmo genético (ver Tabla 5.1). El único

Descripción	Código PDB	nº fragmentos	nº residuos en el modelo	
Péptido V3 IIIB de HIV-1	**1b03**	2	18	
Péptido mutante del extremo N-terminal de la ubiquitina	**1e0q**	2	17	
Péptido MBH12 de 14 residuos RG-KWTY-NG-ITYE-GR	**1k43**	2	14	
Péptido V3 MN	**1niz**	2	14	
Dominio FBP28WW de Mus musculus	**1e0l**	3	23	
Dominio WWIII RNedd4	**1i5h**	3	24	
Dominio WW de Pin1	**1i6c**	3	25	
Dominio Yap65 (mutante L30K)	**1jmq**	3	24	
Dominio WW del factor de *splicing* PRP40 de levadura (fragmento)	**1o6w**	3	25	
Dominios WW3-4 del supresor de deltex (fragmento)	**1tk7**	3	25	
Enzima UBC (E2) del sistema de anclaje de ubiquitina de Caenorhabditis elegans (fragmento)	**1q34**	4	52	

Figura 5.9: Conjunto de proteínas para el estudio de potenciales de enlace de hidrógeno con el número de fragmentos en el modelo, el número de residuos y una representación de su estructura tridimensional.

Parámetros de la minimización	
tamaño de la población	100
nº generaciones por ciclo	1000
p_{cross}	0.5
p_{mut}	0.1
$RMSD_{ij}^{min}$	1. Å

Tabla 5.1: Parámetros del programa evolutivo para la minimización de la energía con potenciales de enlace de hidrógeno.

parámetro que cambia es el número de minimizaciones por ciclo de optimización. Para proteínas de dos fragmentos, cada ciclo de optimización consta de cinco minimizaciones independientes; para las de tres, consta de diez; y para la de cuatro, consta de veinte. Este cambio se justifica porque al aumentar el número de fragmentos aumenta también el número de variables a optimizar. Además, al haber un mayor número de residuos, hay un mayor número de centros de interacción, y por tanto se vuelve más compleja la superficie de energía.

5.4. Resultados de la minimización de la energía

5.4.1. Modelo de Irbäck

En la Tabla 5.2 se muestran los resultados obtenidos con el modelo de Irbäck. Para cada proteína presentamos la energía de la conformación nativa (E_{Nat}), la energía mínima alcanzada en la optimización (E_{Min}) y el valor de $RMSD$, calculado entre carbonos-α, del mínimo con respecto a la nativa en Å. Lo primero que llama la atención en la Tabla es que, para cinco de las once proteínas, la energía de la conformación calculada para la conformación nativa es positiva. Esto supone que, en estos casos, de acuerdo al potencial de Irbäck, el conjunto de los enlaces de hidrógeno es levemente desfavorable para la estabilidad de la estructura nativa. Las proteínas son sistemas mínimamente

$PDBid$	E_{Nat}	E_{Min}	$RMSD$ (Å)
1b03	2.25	-2.27	0.28
1e0q	-1.34	-3.49	0.42
1k43	3.74	-1.93	0.63
1niz	-1.70	-2.07	0.41
1e0l	-0.288	-4.68	0.98
1i5h	-2.28	-4.93	0.64
1i6c	0.501	-5.42	0.76
1jmq	-1.79	-3.29	4.1
1o6w	-2.67	-5.92	0.60
1tk7	1.12	-5.83	0.84
1q34	3.38	-11.82	0.47

Tabla 5.2: Resultados de la optimización con el potencial de Irbäck: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}), ambas en unidades arbitrarias, y valor de $RMSD$ en angstroms del mínimo con respecto a la nativa.

frustrados[22], por lo que es posible, aunque infrecuente, que algunas interacciones se encuentren lejos de su mínimo energético. Esto explica por qué, en una determinada conformación, el valor de la energía de los distintos enlaces de hidrógeno puede variar en un cierto intervalo. Sin embargo, en los casos que ahora analizamos, los valores positivos de energía para el estado nativo se deben más a la parametrización del modelo que a la frustración del sistema. En concreto, el parámetro responsable de los valores positivos de energía es la distancia de equilibrio σ entre los átomos interaccionantes, cuyo valor es 2 Å. Como explicamos para el contacto repulsivo que aparecía en 2gb1, para este valor de σ el punto de corte del potencial de Lennard-Jones con el eje de abscisas es 1.83 Å. Por tanto, para valores inferiores de distancia entre H y O la energía es positiva. En las proteínas que hemos estudiado hay muchos enlaces de hidrógeno con distancias H-O inferiores a 1.83 Å. La menor de ellas es 1.64 Å, lo cual supone una diferencia de sólo 0.2 Å con respecto al punto de corte. Aunque no se trata de una diferencia importante, es suficiente para hacer positiva la energía global de la conformación nativa de la proteína. Podría criticarse el valor propuesto por Irbäck *et al.* para la distancia de equilibrio, pero

se trata del valor habitual observado en enlaces de hidrógeno entre grupos carbonilo y amino del esqueleto de proteínas[1]. Su elección, por tanto, está plenamente justificada. También la forma funcional, rápidamente creciente en la parte repulsiva, es en parte responsable de los valores positivos de energía.

Otro aspecto a destacar en los valores de energía de la Tabla 5.2 es la importante diferencia entre la energía de la conformación nativa y el valor de energía mínima para cada proteína. Estas diferencias se deben en parte a que en las estructuras de mínima energía no aparecen repulsiones como las que encontrábamos en las formas nativas. Pero también para aquellas proteínas sin repulsiones nativas la mejora en el valor de la energía es sustancial. El método de minimización permite alcanzar conformaciones en las que los valores de las distancias r_{ij} y los ángulos α y β (definidos en la Figura 5.1), para pares de centros i y j de distintas hebras que están interaccionando se acercan mucho a sus valores óptimos. En concreto, las distancias tratan de aproximarse lo más posible a 2 Å, y los ángulos α y β se aproximan a la orientación en la que los enlaces C=O y N-H están alineados. Este ajuste hacia los valores de distancias y ángulos óptimos se produce en la medida en que lo permite la rigidez de los fragmentos. Por ello, los valores alcanzados para cada par de centros de interacción ij son siempre valores de consenso sobre la estructura global, que pueden no corresponder al mínimo de u_{ij}.

En la Tabla 5.2 mostramos también los valores de $RMSD$ de las estructuras de mínima energía frente a la nativa. Para todas las proteínas que hemos estudiado, excepto 1jmq, estos valores son inferiores a 1 Å. Salvo en esa excepción —en la que profundizaremos más adelante— el potencial define una superficie de energía que tiene su mínimo muy próximo a la conformación nativa. No obstante, a lo largo del muestreo se encuentran a menudo mínimos locales. Estos mínimos locales corresponden a conformaciones con disposiciones alternativas de las hebras. Por ejemplo, en la Figura 5.10 mostramos la representación de las conformaciones de mínima energía que se van alcanzando a lo

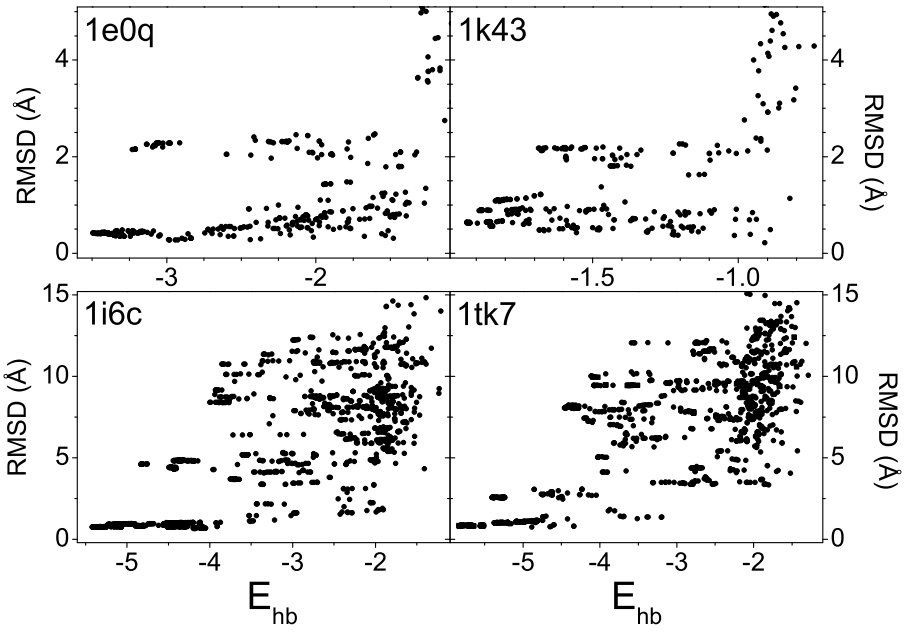


Figura 5.10: Representaciones de energía frente a $RMSD$ con respecto la nativa (en Å) de las mejores conformaciones de varias proteínas —1e0q, 1k43, 1i6c y 1tk7— exploradas con el modelo de Irbäck. En cada panel se muestran juntos los resultados de 5 ejecuciones del programa.

largo de la minimización con el modelo de Irbäck para varias proteínas: 1e0q y 1k43, ambas de dos fragmentos, o 1i6c y 1tk7, de tres. En cada panel, un punto representa una conformación que en alguna etapa de la optimización ha sido la de mínima energía. Cada conformación está definida por sus valores de energía E_{hb} y $RMSD$ frente a la nativa. Para cada una de las proteínas, se muestran juntos los resultados de todas las optimizaciones llevadas a cabo. En todos los casos, la población de puntos con valores más bajos de energía aparece para valores de $RMSD$ muy pequeño. Por tanto, corresponde a una estructura muy parecida a la nativa, como habíamos visto en la Tabla 5.2. En los cuatro paneles se puede apreciar la aparición de otras poblaciones de puntos. Se trata de los mínimos locales que hemos mencionado, es decir, otros posibles ordenamientos de las hebras que son explorados durante el muestreo.

Hemos inspeccionado visualmente la estructura de los mínimos locales. Como he-

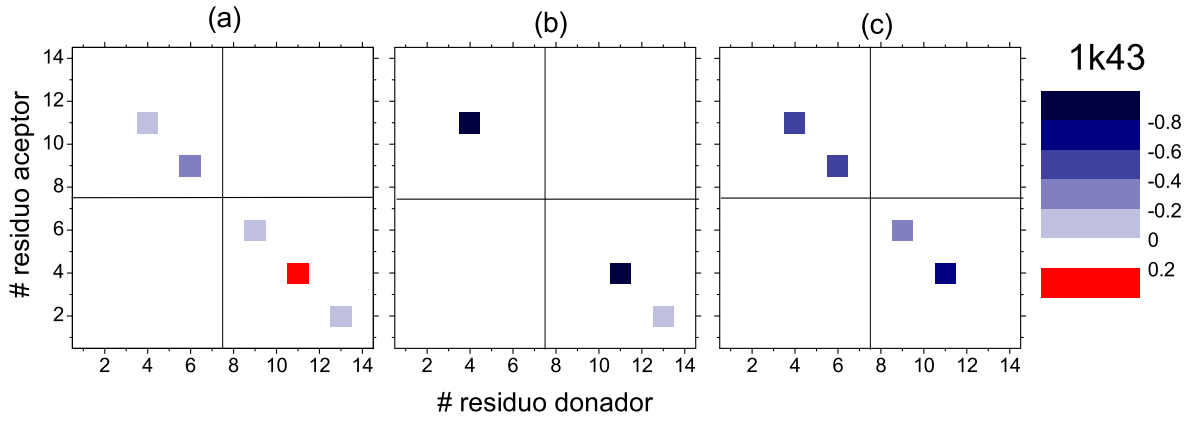


Figura 5.11: Mapas de energía de 1k43 con el modelo de Irbäck. (a) Conformación nativa, (b) mínimo local con $E_{hb} = -1.68$ y $RMSD$= 2.13 Å, y (c) el mínimo absoluto con $E_{hb} = -1.93$ y $RMSD$= 0.63 Å.

mos comentado, corresponden a conformaciones con una disposición alternativa de los fragmentos en que dividimos la proteína. Como en el mínimo global, en una disposición alternativa ya no aparecen —si las había— repulsiones nativas, aun a costa de no satisfacer enlaces de hidrógeno nativos. En la Figura 5.11 mostramos mapas de energía para tres conformaciones de 1k43. El primero de los mapas (Figura 5.11 (a)) corresponde a la conformación nativa. En él vemos que se forman cinco enlaces de hidrógeno, dos entre los residuos 4 y 11 (H_4-O_{11} y H_{11}-O_4), otros dos entre los residuos 6 y 9 (H_6-O_9 y H_9-O_6), y finalmente otro entre el 13 y el 2 (H_{13}-O_2). La interacción H_{11}-O_4, representada en rojo, es desfavorable según el modelo, debido a que la distancia entre centros es 1.64 Å. En el mapa del mínimo local (Figura 5.11 (b)), correspondiente a una conformación con energía $E_{hb} = -1.68$ y $RMSD$=2.13 Å, este contacto repulsivo ya no aparece. Sin embargo, tampoco se forman los enlaces de hidrógeno que se establecían en la estructura nativa entre los residuos 6 y 9. Finalmente, en la conformación optimizada, con energía $E_{hb} = -1.93$ y $RMSD$=0.63 Å (Figura 5.11 (c)) se recuperan los enlaces de hidrógeno nativos, excepto el más débil, H_{13}-O_2.

Cuando estudiamos proteínas con mayor número de fragmentos, aumenta el núme-

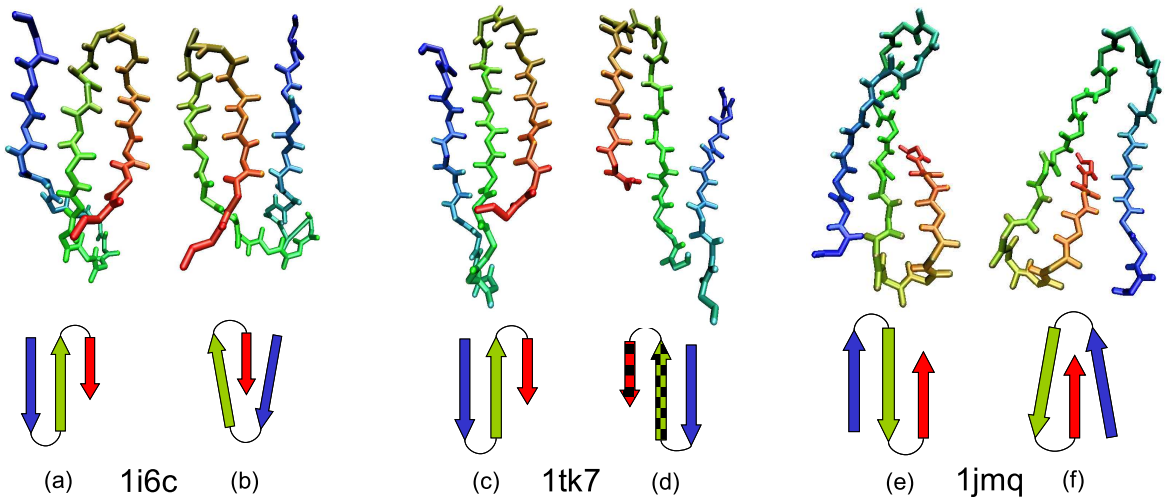


Figura 5.12: Representaciones esquemáticas y mapas topológicos de varias proteínas. 1i6c: (a) conformación nativa, (b) mínimo local; 1tk7: (c) conformación nativa, (d) mínimo local; 1jmq: (e) conformación nativa, (f) mínimo global. El ajedrezado en los mapas topológicos indica que los fragmentos se encuentran dados la vuelta.

ro de conformaciones de baja energía posibles. Así, para 1i6c y 1tk7 aparecen mínimos locales con mayores diferencias estructurales con respecto a la nativa. En la Figura 5.12 mostramos representaciones de la conformación nativa de 1i6c (a) y de un mínimo local encontrado durante la búsqueda (b). En la nativa, las hebras se disponen consecutivamente de acuerdo con la secuencia (1-2-3), mientras que en el mínimo local la hebra $\beta 3$ (en rojo) se encuentra intercalada entre la $\beta 2$ (en verde) y la $\beta 1$ (en azul), de modo que la disposición sería 1-3-2. El caso de la proteína 1tk7 es parecido al de 1i6c (ver Figura 5.12 (c) y (d)). Aquí, en el mínimo local el dominio formado por las hebras $\beta 2$ y $\beta 3$ (en verde y rojo, respectivamente) se encuentra dado la vuelta en torno al eje de la hebra $\beta 1$ (en azul), con respecto a su situación en la nativa.

La única excepción a los excelentes resultados del modelo de Irbäck es la proteína 1jmq, de tres fragmentos (ver Tabla 5.2). El mínimo energético tiene un valor de $RMSD$ frente a la conformación nativa próximo a 4 Å. En la Figura 5.12 representamos el empaquetamiento nativo de las hebras (e), y también el de menor energía (f). En el mínimo se conserva la horquilla formada por las hebras $\beta 2$ y $\beta 3$ (en verde y rojo,

respectivamente) que aparece en la nativa. Pero con respecto a este grupo, la hebra $\beta 1$ (en azul) cambia de posición. Es decir, que para 1jmq el modelo no es capaz de definir el mínimo en la conformación nativa. Aun así, localmente sí se mantiene la disposición de hebras de la estructura nativa.

5.4.2. Modelo de Chen

En la Tabla 5.3 mostramos los valores de energía obtenidos con este modelo para la conformación nativa de cada una de las proteínas. El superíndice que acompaña al valor de la energía nativa en la Tabla es el número de veces que se violan las restricciones de volumen excluido del modelo de Chen. Como hemos comentado, en nuestra implementación de este modelo el diámetro de esferas duras para carbonos-α vale 4.56 Å. Como ya sucedía con 2gb1, en las conformaciones nativas de nuestro conjunto de proteínas esta restricción se quebranta en numerosas ocasiones. De acuerdo con la definición de los autores, para todas las proteínas de la Tabla la energía de la conformación nativa asignada por el potencial sería infinita. En la Tabla, para la conformación nativa, mostramos valores que proceden de la parte atractiva del potencial, en ausencia de repulsiones. Así, podemos compararlos con los valores de energía mínima alcanzados en la optimización. Dado el gran número de repulsiones en las estructuras nativas, tenemos que pensar que el diámetro de esferas duras propuesto tiene un valor demasiado elevado, incompatible con las láminas β reales, como ya apuntaba el resultado obtenido para 2gb1.

A partir de nuestros experimentos de minimización se puede comprobar si este problema impide o no alcanzar el mínimo energético en una conformación vecina a la nativa. En la Tabla 5.3 se muestran los valores de energía mínima y de $RMSD$ frente a la nativa de la correspondiente conformación para cada una de las proteínas. En muchos casos los valores de $RMSD$ son muy bajos, a menudo inferiores a 1 Å. Por tanto, el potencial de Chen consigue con razonable éxito definir el mínimo de energía

$PDBid$	E_{Nat}	E_{Min}	$RMSD$ (Å)
1b03	-0.576^3	-0.547	1.34
1e0q	-0.558	-0.964	1.85
1k43	-0.674^2	-0.788	0.49
1niz	-0.662^2	-0.757	0.58
1e0l	-1.11^5	-1.40	1.31
1i5h	-1.18^3	-1.45	1.02
1i6c	-1.10^4	-1.36	1.92
1jmq	-1.21^8	-1.32	0.81
1o6w	-1.21^3	-1.38	1.49
1tk7	-1.57^6	-1.94	0.96
1q34	-2.87^5	-2.67	0.91

Tabla 5.3: Resultados de la optimización con el potencial de Chen: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}) y valor de $RMSD$ en angstroms del mínimo con respecto a la nativa. El superíndice sobre los valores de E_{Nat} indica el número de veces que se incumplen las restricciones de volumen excluido.

en las vecindades de la conformación nativa, a pesar de las repulsiones que aparecen en esta. Para alcanzar estos mínimos energéticos, se evitan los contactos repulsivos que aparecen en la estructura nativa y tratan de aproximarse las distancias H$'$-O$'$ entre centros interaccionantes a su valor de equilibrio.

Un ejemplo del buen funcionamiento del modelo es el caso de 1b03, una de las horquillas β que hemos estudiado. En la Figura 5.13 mostramos los mapas de enlaces de hidrógeno del modelo de Chen para esta proteína. Estos mapas corresponden a la conformación nativa y a la de menor energía tras la minimización. Las interacciones más fuertes en ambas conformaciones son las mismas en la optimizada y en la nativa, las que se establecen entre los residuos 2-17 y 4-15. Vemos, en todo caso, que en la conformación de mínima energía algunas interacciones se han debilitado. Esto determina el valor de la energía mínima, menor en valor absoluto que el término atractivo de la nativa. Sin embargo, no se detectan las repulsiones que aparecían en la nativa, representados en rojo en la Figura 5.13 (a). El valor de $RMSD$ (1.34 Å) no es de los más bajos que hemos

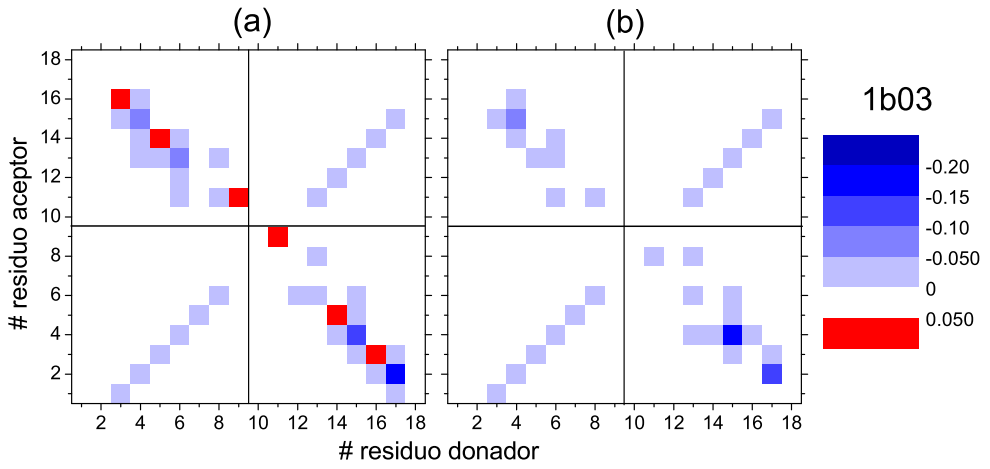


Figura 5.13: Mapas de energía de 1b03 con el modelo de Chen. (a) Conformación nativa, y (b) mínimo absoluto con $E_{hb} = -0.547$ y $RMSD$= 1.34Å.

obtenido con este potencial. Aun así, la conformación de mínima energía sigue siendo próxima a la nativa.

Como observábamos con el modelo de Irbäck, el de Chen propicia el muestreo, a lo largo de la optimización, de mínimos locales que corresponden a conformaciones alternativas a la nativa. Esto no impide que se encuentre el mínimo en el entorno de la nativa para todas las proteínas estudiadas. Hay un caso en el que aparecen dos estados energéticamente degenerados, la proteína 1e0q. En la Figura 5.14 representamos energía frente a $RMSD$ de las estructuras de mínima energía calculadas a lo largo de la optimización para esta proteína. Vemos que se localizan dos mínimos con un valor muy parecido de energía, pero distinto valor de $RMSD$ frente a la nativa. Uno de ellos aparece con $RMSD$ de 1.85 Å, y corresponde a una disposición alternativa de las hebras. En ella, una de las hebras se encuentra girada sobre su propio eje con respecto a su posición en la conformación nativa. El otro mínimo tiene $RMSD$ de 1.11 Å y es una conformación muy parecida a la nativa. Esta proteína es la única para la que no se quebranta en ninguna ocasión la restricción de volumen excluido. Además, de todos los casos de dos hebras, es la que tiene una disposición más planar. La aparición

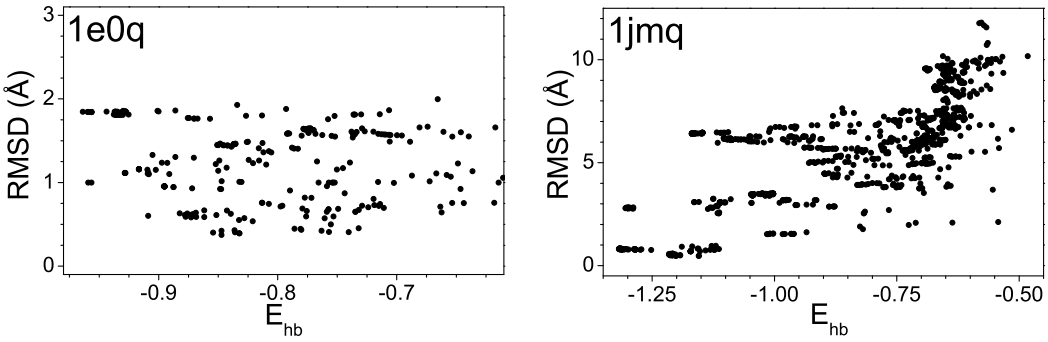


Figura 5.14: Resultados de la energía frente al $RMSD$ con respecto a la nativa (en Å) de las mejores conformaciones de 1e0q y 1jmq con el modelo de Chen.

de degeneración en este caso puede indicar una tendencia del modelo a la degeneración entre empaquetamientos con disposiciones alternativas de las hebras, especialmente para estructuras menos determinadas por la topología de los fragmentos rígidos.

También en el caso de 1jmq observamos este tipo de degeneración. En la representacion de la energía frente a $RMSD$ de las conformaciones de mínima energía a lo largo de la optimización que mostramos en la Figura 5.14, se observa que hay dos mínimos con la misma energía y distinto $RMSD$. En la Figura 5.15 mostramos los mapas de energía para la estructura nativa, el mínimo nativo y el mínimo alternativo. Ni en el mínimo de $RMSD$=0.81 Å ni en el de 2.79 Å aparecen las interacciones desfavorables entre carbonos-α de la estructura nativa. Por otro lado, los enlaces de hidrógeno más fuertes del mínimo de $RMSD$=0.81 Å (Figura 5.15 (c)) se corresponden muy bien con las de la estructura nativa (Figura 5.15 (a)). En cambio, en el mapa del mínimo alternativo la tercera hebra se encuentra girada con respecto a su posición nativa Esto ese traduce en un cambio en el registro y, por tanto, en el patrón de contactos entre las hebras $\beta 2$ y $\beta 3$ (Figura 5.15 (b)).

Como hemos comentado, el modelo ofrece muy buenos resultados para casi todas las proteínas. Las pequeñas diferencias entre la conformación nativa de una proteína y su mínimo energético proceden a menudo de la falta de dependencia orientacional en

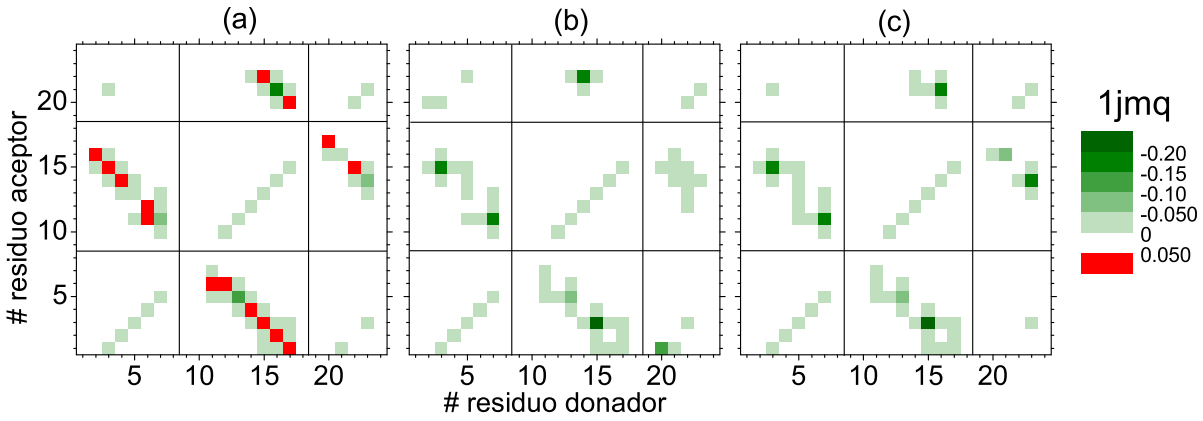


Figura 5.15: Mapas de energía de 1jmq con el potencial de Chen: (a) conformación nativa, (b) mínimo alternativo, con $E = -1.30$ y $RMSD = 2.79$ Å y (c) mínimo nativo, con $E = -1.32$ y $RMSD = 0.81$ Å.

el potencial. Por ejemplo, el mínimo global para la horquilla β 1k43 tiene un valor de $RMSD$ de tan solo 0.49 Å, es decir, que la conformación nativa y la optimizada son muy parecidas. Hemos representado su traza de carbonos-α en la Figura 5.16, en azul la nativa y en rojo, la optimizada. Además, para mayor claridad, hemos representado los planos que contienen a las hebras (como flechas amarillas para la nativa y rojas para la optimizada). La conformación de mínima energía puede obtenerse por una transformación de la nativa que consiste en una rotación del plano de la hebra $\beta 2$ con respecto al plano de la hebra $\beta 1$. Este comportamiento, que encontramos también para casos con un mayor número de fragmentos, es resultado de la falta de dependencia angular en la función de energía.

En algunos casos de proteínas de tres fragmentos, como 1e0l, se definen mínimos locales no sólo en disposiciones alternas de las hebras en la lámina, sino también en "bundles" o "haces" de hebras. En la Figura 5.17 mostramos la conformación nativa (a) y el mínimo local (b) de 1e0l. En el haz de hebras, los fragmentos no están formando un plano como en la lámina nativa, sino que cada uno de ellos interacciona fuertemente con los otros dos. De tal manera que el potencial de Chen tiene aquí un efecto de colapso

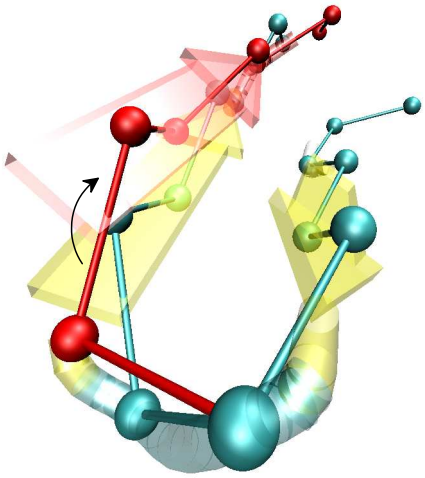


Figura 5.16: Rotación de una hebra de 1k43 que permite transformar la conformación nativa (en azul) en la optimizada (en rojo) con el modelo de Chen.

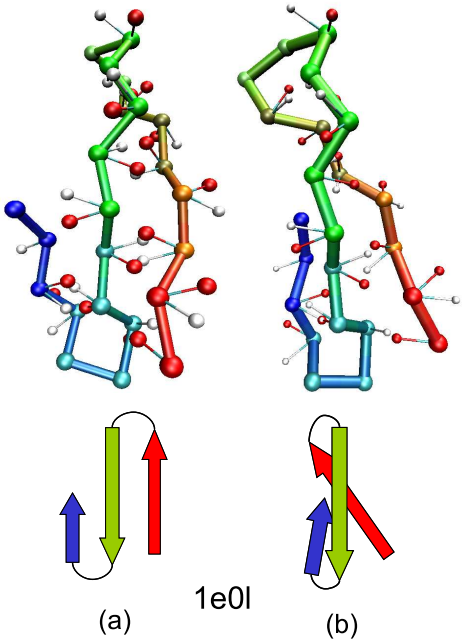


Figura 5.17: Representaciones esquemáticas y mapas topológicos para 1e0l. (a) Conformación nativa, y (b) mínimo local de $E = -1.33$ y $RMSD$= 2.20 Å.

inespecífico. Este tipo de estructuras se pueden formar debido, de nuevo, a la falta de una direccionalidad en la definición del enlace de hidrógeno del modelo. Al formarse un haz, se maximiza el número de interacciones por residuo. Sin embargo, el modelo evita el peligro de que se estabilicen en exceso estructuras alternativas, como la descrita para 1jmq o haces de hebras como el de 1e0l, gracias a la repulsión de volumen excluido de diámetro grande.

Comentamos finalmente dos aspectos especialmente importantes para que el modelo de Chen sea tan eficaz como se muestra en estos resultados. El primero de ellos se refiere a los residuos que forman el lazo que une las hebras. Como hemos comentado en la introducción, el modelo de Chen parte de un homopolímero cuyos monómeros pueden formar enlaces de hidrógeno. En el caso del homopolímero, todos los monómeros son iguales, y el modelo para este tipo de molécula tiene su mínimo energético en

conformaciones helicoidales. Después, los autores transforman el homopolímero en heteropolímero, haciendo que una serie de residuos hacia la mitad de una cadena no sean capaces de interaccionar. Así se favorece la aparición de estructuras de tipo lámina. En nuestra implementación de su modelo partimos de las hebras como fragmentos rígidos, lo cual supone una clara ventaja para la formación de láminas. Aun así hemos observado que, si no se anula la posibilidad de interaccionar de los residuos entre fragmentos, los resultados empeoran notablemente. Como hemos mostrado para 2gb1, la reconstrucción de posiciones de hidrógenos y oxígenos virtuales, partiendo de la conformación nativa, es más incorrecta en esas regiones. Por ello, si se consideran sus interacciones, se estabilizan artificialmente mínimos locales como los que hemos descrito.

El segundo aspecto que queremos señalar está relacionado con el método de reconstrucción de posiciones de átomos virtuales. Hay dos residuos de la cadena de los cuales sólo se puede reconstruir un centro de interacción. Estos residuos son el primero, del que sólo podemos calcular la posición del oxígeno virtual, y el penúltimo, del que se calcula únicamente la posición del hidrógeno. Además, del último residuo no se puede reconstruir ninguno de los átomos virtuales. En los resultados que hemos descrito hasta ahora, hemos considerado tanto el primer oxígeno como el penúltimo hidrógeno virtuales, como hacían los autores en sus trabajos. Hemos comprobado qué sucedería en ausencia de estas interacciones, es decir, si sólo considerásemos aquellos aminoácidos cuyos centros de interacción pueden ser reconstruidos completamente. En este caso, los resultados son notablemente peores. Por tanto, la inclusión de estos centros de interacción al principio y al final de la cadena es crítica para que se defina el mínimo energético en la conformación nativa.

5.4.3. Modelo de Kolinski

En la Tabla 5.4 presentamos los resultados obtenidos con este potencial para el conjunto de proteínas estudiadas. Los valores de energía para la conformación nativa (E_{Nat}) para todas las proteínas son negativos y siguen una tendencia creciente con el tamaño de la proteína. Para cuatro de las proteínas se infringen las restricciones de volumen excluido (el número de violaciones se incluye como un superíndice en la Tabla 5.4). Para las cuatro proteínas, la restricción que se quebranta es el volumen excluido entre el carbono-α de un residuo y un átomo virtual del modelo (PB). En el modelo CABS, los átomos PB se sitúan entre dos carbonos-α contiguos. Su posición no coincide con ninguna posición atómica real de la proteína, sino con el punto medio de un enlace peptídico. Por tanto, si pensamos en el esqueleto de una proteína real, el átomo PB se colocaría entre un grupo donador y otro aceptor de hidrógeno. Esto justifica que se consideren estos centros para definir los enlaces de hidrógeno en el modelo CABS. Pero también explica el porqué de esas violaciones del diámetro de volumen excluido, $\phi_{C^\alpha - PB}$. En esa región, los esqueletos de dos fragmentos de una proteína que están formando enlace de hidrógeno se aproximan mucho debido a la interacción. El diámetro de esferas duras es suficientemente pequeño para la mayoría de casos en proteínas reales. Por eso normalmente no se quebranta esta restricción. Pero debido a entornos químicos particulares, pueden encontrarse distancias C_α-PB inferiores a 4.2 Å.

Como indican los valores de $RMSD$ (ver Tabla 5.4), para la gran mayoría de proteínas estudiadas el mínimo de menor energía corresponde a una conformación próxima a la nativa. En cuanto a los valores de energía, encontramos diferencias notables entre los valores de la nativa (E_{Nat}) y la optimizada (E_{Min}). Estas diferencias se deben fundamentalmente a dos factores. El primero de ellos es que, en la mayoría de casos, en la estructura de mínima energía los fragmentos se encuentran más próximos que en la nativa. Según la definición original, se calcula la energía de interacción u_{ij} sólo cuando la

$PDBid$	E_{Nat}	E_{Min}	$RMSD$ (Å)
1b03	-13.0[1]	-25.2	1.31
1e0q	-9.43	-30.9	1.19
1k43	-15.4	-28.3	0.64
1niz	-16.9	-28.7	0.45
1e0l	-27.7	-50.4	1.44
1i5h	-21.9	-46.9	1.23
1i6c	-23.2	-47.2	1.42
1jmq	-22.3[6]	-44.9	3.40
1o6w	-29.6	-49.5	1.04
1tk7	-40.1[3]	-68.1	0.89
1q34	-50.9[1]	-77.5	2.79

Tabla 5.4: Resultados de la optimización con el modelo de Kolinski: energía de la conformación nativa (E_{Nat}), mínimo energético alcanzado (E_{Min}) y $RMSD$ en angstroms del mínimo con respecto a la nativa. El superíndice sobre algunos valores de E_{Nat} indica el número de veces que se incumplen las restricciones de volumen excluido.

distancia entre centros es inferior a 6.1 Å. Además, los dos términos de la Ecuación (5.11) se minimizan para distancias pequeñas entre las hebras. En el caso del término u_{ij}^{h}, esto es así por las distancias r_{pp} y r_{qq} entre átomos virtuales enfrentados, vecinos a los carbonos-α interaccionantes. En el caso del término u_{ij}^{γ}, por la distancia $\mid \mathbf{r_i} - \mathbf{r_j} \mid$ entre carbonos-α. Al ser más pequeñas estas distancias en la conformación optimizada que en la nativa, la energía global debida a los enlaces de hidrógeno disminuye.

En la Figura 5.18 mostramos los mapas de distancias y energías para 1e0q con el modelo de Kolinski. A diferencia de los que mostrábamos para los modelos de Irbäck y Chen, aquí en cada mapa se representan los valores para la conformación nativa, en el cuadrante superior izquierdo, y la optimizada, en el inferior derecho. En el mapa de distancias (Figura 5.18 (a)) se representan las distancias entre carbonos-α para cada par de residuos. Se puede observar que la trama de contactos para la conformación optimizada es más tupida que en la nativa. Esto se traduce en la aparición de nuevas interacciones en la conformación optimizada, y en una mayor intensidad en las interacciones que

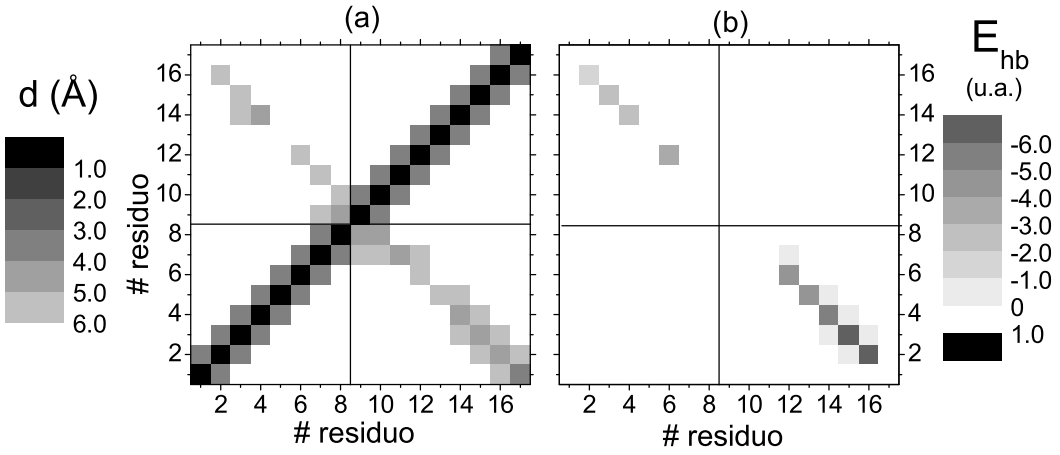


Figura 5.18: Mapas de distancias (a) y energías (b) de 1e0q con el potencial de Kolinski: en ambos se representan los valores de la conformación nativa —cuadrante superior izquierdo— y la optimizada —cuadrante inferior derecho.

aparecían en la nativa, como se puede ver en el mapa de energías (Figura 5.18 (b)). Por ejemplo, todos los contactos que aparecen coloreados en gris más claro en el cuadrante inferior derecho de este mapa, exceptuando el 3-14, corresponden a residuos que en la conformación nativa no satisfacen la restricción de proximidad y sí lo hacen en la optimizada. Además, en este mismo cuadrante vemos que aparece una nueva interacción en la propia diagonal, el 5-13, cuyos centros de interacción, en la nativa, también están más alejados que la distancia de corte.

En el mapa de energías (Figura 5.18 (b)), en la región correspondiente a la conformación optimizada aparecen interacciones débiles fuera de la diagonal. Su aparición se debe a que el modelo permite que se calcule la energía para pares de residuos no enfrentados entre sí. Se trata, por tanto, de residuos que no estarían formando enlaces de hidrógeno reales. Estos residuos no pueden satisfacer la última de las restricciones, pero sí pueden cumplir las tres primeras, por lo cual $\delta^{\gamma} = 1$. El valor mínimo de estas interacciones es -0.438 y su contribución es, en todo caso, pequeña. Estas interacciones contribuyen poco a la energía global de la conformación optimizada, pero son producto de las simplificaciones del modelo.

Como hemos comentado, la primera de las razones por las que disminuye tanto la energía entre la conformación nativa y la optimizada es la disminución de la distancia entre fragmentos. La segunda es que, de acuerdo con el potencial, en la conformación de mínima energía las hebras están mejor orientadas unas con respecto a otras que en la nativa. Varios términos de la compleja función de energía están relacionados con la orientación. En primer lugar, el módulo del término denominado u_{ij}^{γ} depende de la orientación, por el producto escalar que encontramos en la Ecuación (5.13). Además, para determinados pares de centros de interacción, un cambio en la orientación permite que se cumplan en la estructura optimizada restricciones que se quebrantan en la estructura nativa. Por ejemplo, en la conformación optimizada de 1e0q (Figura 5.18 (b)) vemos que aparece la interacción entre los residuos 3 y 14, que no aparece en la nativa (Figura 5.18 (a)). En la conformación nativa los vectores $\mathbf{h}_3$ y $\mathbf{h}_{14}$ forman un ángulo de 53.8°, superior a los 40° que se establecen como máximo en la segunda restricción. En la conformación optimizada, las hebras están levemente rotadas con respecto a su orientación en la nativa y encontramos que el ángulo que forman $\mathbf{h}_3$ y $\mathbf{h}_{14}$ es de 30.8°, inferior al ángulo umbral. Debido al cumplimiento de las tres primeras condiciones se calcula el término $u_{3,14}^{\gamma}$ correspondiente a la interacción, que recibe un valor de -0.419.

La componente orientacional también afecta al cumplimiento de la cuarta de las restricciones del modelo, la que permite que se calcule el término u_{ij}^{h} de la energía. En el mapa de energías de 1e0q que mostramos en la Figura 5.18 (b), las interacciones de la diagonal en la conformación optimizada son mucho más intensas que en la nativa. Esto se debe a que la cuarta restricción se satisface más veces. En la nativa, para la mayoría de pares de residuos interaccionantes (2-16, 3-15 y 4-14) esta restricción sólo se cumple para uno de los dos vectores $\mathbf{h}$, y por tanto δ^h=1. Únicamente para el par 6-12 se cumple con los dos vectores $\mathbf{h}$ (δ^h=2). En cambio, y debido a los ajustes orientacionales con respecto a la nativa, en la optimizada δ^h=2 para todos los pares de residuos que

hemos mencionado, y además δ^h=1 para los pares 5-13 y 7-11. Esto hace que la energía de los correspondientes términos u_{ij}^h sea muy superior en la conformación optimizada que en la nativa.

Como hemos visto en el caso de 2gb1, la sensibilidad de las restricciones hace que pequeñas variaciones en la estructura nativa supongan importantes diferencias en la energía. Debido al gran número de restricciones, el muestreo del espacio conformacional usando fragmentos rígidos impone una limitación más importante para la evaluación de este potencial que en el de Irbäck o el de Chen. Un cambio en la orientación de los vectores correspondientes a un centro de interacción va ligado a cambios en la orientación de los demás centros de este mismo fragmento. Sin embargo, para la mayoría de proteínas el mínimo energético se define correctamente. Por tanto, podemos considerar que las grandes fluctuaciones en energía entre la conformación nativa y la optimizada no son especialmente problemáticas.

Un último aspecto que queremos destacar en el modelo de Kolinski está relacionado con la recomendación de los autores de no considerar las interacciones $(i, i+4)$[103]. En los cálculos cuyos resultados hemos comentado hasta ahora, estas interacciones no han sido tenidas en cuenta, siguiendo las indicaciones de los trabajos originales. Como comprobación, hemos llevado a cabo las mismas optimizaciones incluyéndolas. En la mayoría de los casos, las diferencias entre resultados no son significativas. La diferencia en $RMSD$ frente a la nativa de los mínimos con y sin estas interacciones es aproximadamente de 0.1 Å en la mayoría de los casos. Sin embargo, hay proteínas para las que sí que se observa un cambio muy importante. En la Figura 5.19 (a) mostramos la representación de la energía de las mejores conformaciones frente a $RMSD$ a lo largo de la optimización, con y sin las interacciones $(i, i+4)$. Si no tenemos en cuenta estas interacciones, aparece un solo mínimo global, con $RMSD$ frente a la nativa próximo a cero (ver Tabla 5.4). Por el contrario, si se consideran estas interacciones, aparece otro mínimo en una conformación

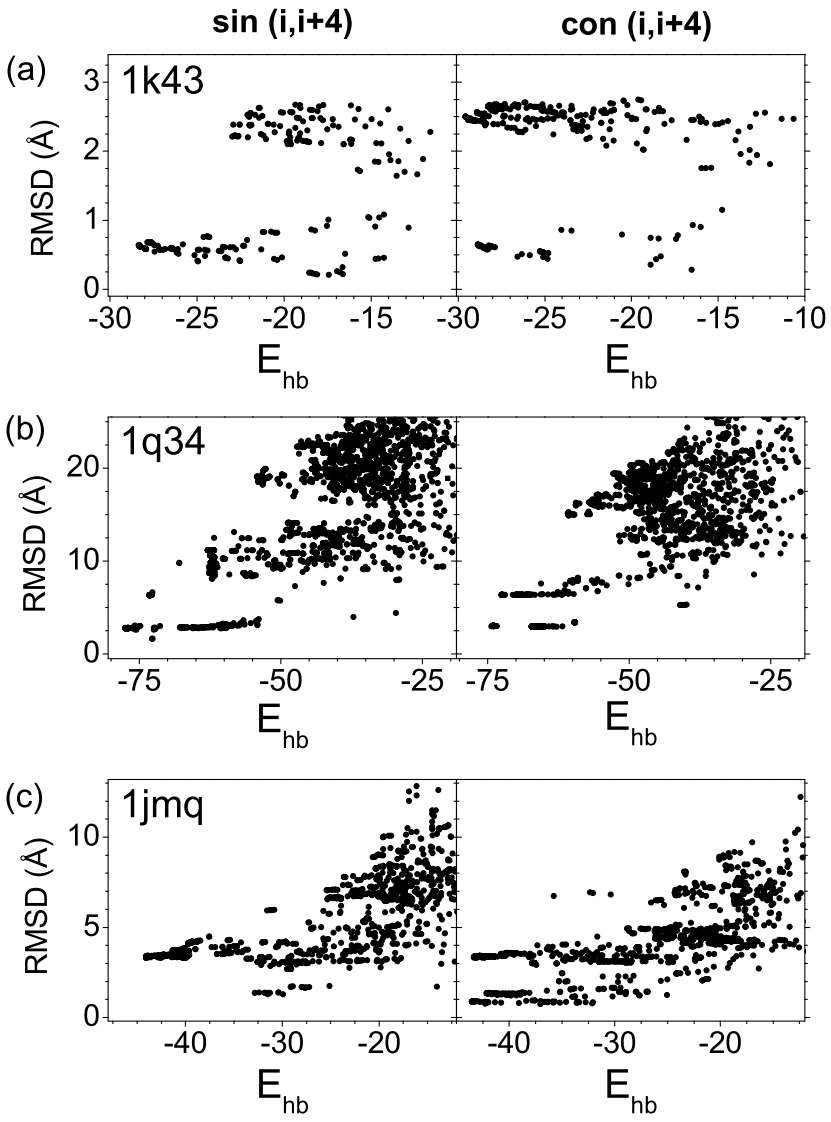


Figura 5.19: Representación del camino de la minimización para 1k43 y 1jmq con el potencial de Kolinski, considerando y sin considerar las interacciones entre residuos $(i, i+4)$.

en la que se forma un lazo grueso, con $RMSD$ próximo a 2 Å. El caso de 1q34, cuyos resultados representamos en la Figura 5.19 (b), es similar. También para esta proteína de cuatro fragmentos en el modelo, al considerar las interacciones $(i, i+4)$ aparece un doble mínimo. En este caso, el segundo mínimo aparece con $RMSD$ mayor de 6 Å, en el que el lazo que une las hebras $\beta 3$ y $\beta 4$ adquiere una disposición semejante a una hélice. Ambos ejemplos sirven para apoyar la opción sugerida por los autores de no considerar estos enlaces de hidrógeno en el modelo.

Sin embargo, hemos encontrado también un caso en el que esta contribución puede ser crucial. Se trata de la proteína 1jmq, que dividimos en tres fragmentos, y cuyos resultados mostramos en la Figura 5.19 (c). En este caso, el mínimo que aparece en ausencia de las interacciones $(i, i+4)$ tiene $RMSD$=3.4 Å con respecto a la estructura nativa. Esta conformación corresponde a un ordenamiento alternativo de las hebras. Pero cuando se consideran estas interacciones aparece un segundo mínimo, degenerado en energía, muy próximo a la estructura nativa. Por tanto, en este caso son necesarias las interacciones $(i, i+4)$ para definir correctamente la superficie de energía. El giro β entre las hebras $\beta 1$ y $\beta 2$ en 1jmq pertenece a la clase II$'$ de acuerdo con la clasificación de Wilmot y Thornton[169,170]. Estos giros son levemente más anchos que los de tipo I, los más abundantes en proteínas. Para estabilizar este tipo de estructura distorsionada hacen falta las interacciones $(i, i+4)$ del modelo. Dado que la aparición de este tipo de giros en estructuras nativas de proteínas es infrecuente, no es un gran problema que el modelo de Kolinski no las considere.

5.5. Resumen del Capítulo y conclusiones

Hemos dedicado este capítulo a nuestro estudio de tres modelos para el enlace de hidrógeno en el esqueleto de proteínas, desarrollados en los grupos de Irbäck[102], Chen[158]

y Kolinski[103]. Como hemos comentado, prácticamente todos los residuos en una proteína forman enlaces de hidrógeno a través de átomos del esqueleto. Esto permite a las proteínas adquirir los tipos más importantes de estructura secundaria, la hélice α y la lámina β. Los tres modelos han sido utilizados anteriormente en estudios en los que la contribución de enlace de hidrógeno era sólo uno entre varios términos utilizados para el cálculo de la energía en simulaciones del plegamiento. Los modelos que hemos seleccionado no utilizan una aproximación estadística para reproducir las interacciones reales, a diferencia de los potenciales hidrófobos que evaluamos en el capítulo anterior. Por el contrario, parten de una representación más o menos simplificada de la proteína y utilizan para el cálculo de la energía funciones sencillas, con un pequeño número de parámetros. Es especialmente relevante en nuestra evaluación que los tres modelos tienen características muy distintas entre sí, tanto en la representación de la proteína como en el potencial para las interacciones. Esto nos permite comprender qué se sacrifica al escoger representaciones simplificadas para las interacciones de la proteína frente a otras más detalladas.

El primero de los modelos analizados es el del grupo de Irbäck. En este modelo el esqueleto de la proteína se representa con resolución atómica, es decir, se consideran cinco átomos por residuo. La función de potencial es dependiente de la distancia entre átomos y de la orientación de los grupos interaccionantes. Este modelo se asemeja mucho a cómo se considera la interacción por enlace de hidrógeno en campos de fuerza como AMBER[156] y CHARMM[157]. Este parecido nos autoriza para utilizar el modelo de Irbäck como referencia para la comparación con los modelos de Chen y Kolinski, mucho más simplicados.

En el modelo de Chen se utilizan únicamente las coordenadas reales de los carbonos-α de la proteína. A partir de ellas se generan los dos centros de interacción de enlace de hidrógeno de cada residuo: los átomos virtuales H$'$ y O$'$. Por tanto, a pesar de ser un mo-

delo de grano grueso, refleja una propiedad muy importante de los enlaces de hidrógeno: la disposición de los centros de interacción a un lado u otro de la traza de carbonos-α. El potencial depende sólo de la distancia entre estos centros. Casi tan importante como este término es una contribución de volumen excluido que restringe mucho las estructuras accesibles en el espacio conformacional.

El tercero de los modelos, el de Kolinski, comparte con el de Chen que de la proteína real se toman únicamente las coordenadas de los carbonos-α. A partir de estas posiciones se calculan las de los átomos virtuales PB, localizados entre pares de carbonos-α sucesivos en la secuencia. En este modelo no se reconstruyen los centros de interacción de cada residuo como sucede en el de Chen, sino que mediante una serie de restricciones y dos términos para el potencial, busca reproducir el efecto del enlace de hidrógeno.

Hemos desdoblado nuestro estudio de los modelos para el enlace de hidrógeno en dos partes. En primer lugar, hemos estudiado cómo reproducen los enlaces de hidrógeno sobre la conformación nativa de la proteína 2gb1. En esta estructura aparecen regiones con estructura secundaria de tipo hélice α y lámina β. Como era de esperar, el potencial de Irbäck reproduce los patrones de interacción con gran eficacia, tanto para hélices como para láminas. Además, hemos detectado los problemas de una parametrización sencilla al encontrar una interacción repulsiva en el mapa de energías (ver Figura 5.7). Este contacto repulsivo se debe al valor seleccionado para la distancia de equilibrio en el enlace de hidrógeno, σ=2 Å, en la Ecuación (5.1). Esta repulsión pone de manifiesto la dificultad de lograr un compromiso satisfactorio en el diseño de cualquier potencial simplificado. Con una función de energía con un mínimo más ancho no encontraríamos este problema, pero también sería menos específico.

En el modelo de Chen nos hemos fijado en primer lugar en cómo se representan los centros de interacción virtuales. Una vez reconstruidas sus posiciones a partir de las de los carbonos-α de la conformación nativa, hemos calculado el valor de $RMSD$ de los

átomos de hidrógeno y oxígeno virtuales con respecto a los centros reales. Por diseño del modelo, sabíamos que debía haber una diferencia de aproximadamente 1 Å para cada tipo de centro. Con nuestro cálculo hemos corroborado esta expectativa. Además, hemos podido ver cómo la contribución al valor de $RMSD$ no es homogénea a lo largo de la secuencia (ver Figura 5.8). Las posiciones de los centros de interacción de residuos que se encuentran en lazos que unen elementos de estructura secundaria están mucho peor representados que los que están en hélices o láminas. Hemos mostrado que en el modelo esto no constituye un problema. Debido a la existencia de unos residuos bisagra en la definición original del modelo aparecen hélices y láminas en el mínimo energético para una cadena polipeptídica. En los trabajos originales de Chen e Imamura[158,160], estos residuos bisagra no forman enlaces de hidrógeno. Los residuos para los que el método de reconstrucción no funciona bien corresponden precisamente a los lazos entre fragmentos con estructura secundaria bien definida, es decir, a lo que en el modelo original serían residuos bisagra.

Hay que destacar también que con el modelo de Chen hemos encontrado sistemáticamente repulsiones en la conformación nativa de 2gb1 (ver Figura 5.7). Los autores situan el potencial de esferas rígidas a una distancia tal que, en un muestreo del espacio conformacional de la proteína, las hebras β no podrían encontrarse tan próximas entre sí como en la conformación nativa. En cuanto a las interacciones favorables que aparecen en el mapa, vemos que con el potencial de Chen las interacciones más fuertes son los enlaces de hidrógeno nativos. Además, aparecen muchas otras contribuciones marginales fuera de las diagonales de contactos, que contribuyen a la estabilidad de la estructura. Estas contribuciones se deben a que en la función de energía falta una dependencia explícita con la orientación.

Con el modelo de Kolinski hemos visto que las interacciones se representan de una manera muy distinta a las de Irbäck y Chen. En el mapa de energías se definen muy

bien las diagonales de enlaces de hidrógeno que estabilizan los elementos de estructura secundaria de 2gb1 (ver Figura 5.7). Aquí no aparecen los contactos fuera de la diagonal que sí veíamos con el modelo de Chen, por lo que podemos decir que el modelo de Kolinski es más específico. Pero en este modelo no se definen centros donadores y centros aceptores de enlace de hidrógeno, que se puedan disponer a uno u otro lado de la traza de carbonos-α. Por tanto, todos los residuos pueden formar enlaces de hidrógeno con sus vecinos en la estructura terciaria, lo que lo hace menos preciso que el de Chen. Así por ejemplo, en láminas β no se obtiene el patrón de interacciones propio de los enlaces de hidrógeno de estos elementos de estructura secundaria de proteínas. Por otra parte, con el modelo de Kolinski hemos podido observar el efecto de las severas restricciones que se imponen al cálculo de la energía de enlace de hidrógeno. Por ejemplo, en hebras enfrentadas que forman parte de la lámina β de 2gb1, en ocasiones sucede que dos residuos i y j que forman enlaces de hidrógeno nativos no cumplen alguna de las restricciones. Sin embargo, sí se calculan contribuciones que no corresponden a interacciones nativas. Esto pone de manifiesto la intención de los autores de no reproducir uno a uno los enlaces de hidrógeno de la proteína, sino su efecto global en la estructura.

La segunda parte del estudio ha consistido en una serie de experimentos de minimización de la energía que hemos llevado a cabo con nuestro algoritmo de minimización. Para ello, hemos utilizado un conjunto de proteínas con estructura de tipo lámina β, formadas por hebras unidas por enlaces de hidrógeno. Cada una de las proteínas se considera como un conjunto de hebras β cuyos grados de libertad internos quedan congelados. Así, el muestreo conformacional se realiza entre los posibles empaquetamientos de los fragmentos, como hemos venido haciendo en las distintas partes de esta Tesis. El requisito que imponemos a los modelos es que el empaquetamiento de fragmentos para el que se defina el mínimo energético corresponda a la conformación nativa de la proteína.

El más eficiente de los modelos es el de Irbäck, en concordancia con su mayor resolución. En la mayoría de casos el mínimo se encuentra en una conformación muy parecida a la nativa, con $RMSD$ a menudo menor que 1 Å. Hemos comprobado, asímismo, que las repulsiones que localizamos en la conformación nativa de varias proteínas (ver Tabla 5.2) no impiden la buena definición del mínimo. Las desviaciones con respecto a las interacciones de la conformación nativa son en general muy pequeñas. La única excepción es la proteína 1jmq, dividida para el muestreo conformacional en tres fragmentos, para la que no aparece un mínimo en la conformación nativa. Aun así, en la estructura optimizada hay un dominio formado por dos de las hebras en el que se mantienen las interacciones nativas.

En el caso del potencial de Chen los resultados de los experimentos de minimización también son muy buenos, aunque acaso no tanto como en el modelo de referencia. Como para la proteína 2gb1, en las conformaciones nativas de las proteínas de tipo lámina β, hemos observado un gran número de repulsiones nativas entre carbonos-α por el diámetro de volumen excluido. En general, estas repulsiones no impiden que en muchos casos la estructura optimizada se encuentre muy próxima a la nativa, a menudo con valores de $RMSD$ inferiores a 1 Å. Es más, la definición del volumen excluido permite que el modelo sea capaz de discernir entre empaquetamientos de tipo lámina y de tipo haz. Así, la suma del potencial atractivo de enlace de hidrógeno entre centros virtuales y la contribución de esferas blandas hace que no sea necesaria la dependencia orientacional explícita.

Finalmente, el modelo de Kolinski también ofrece buenos resultados, comparables a los del modelo de Chen. La problemática de las restricciones que apuntaba el estudio de la conformación nativa de 2gb1 aparece también en las minimizaciones energéticas. Así, para varias proteínas, entre la conformación nativa y la optimizada hay pocas diferencias en estructura y una gran variación en la energía. Esta sensibilidad a pequeños

cambios puede estar acentuada por nuestro tipo de búsqueda conformacional, limitada a empaquetamientos de fragmentos rígidos.

La proteína con la que peor funciona el modelo de Kolinski es 1jmq, para la que el mínimo se define en un empaquetamiento alternativo de fragmentos. Este mal funcionamiento se debe a una peculiaridad de la estructura nativa de 1jmq, en la que aparece un giro de tipo II′, infrecuente en proteínas. Estos giros son más gruesos de lo habitual, y para que el modelo sea capaz de reproducirlo haría falta considerar las interacciones $(i, i+4)$, que los autores del potencial no incluyen en los trabajos originales[103]. Este caso particular puede servir de ejemplo de las desventajas de los modelos simplificados frente a los de resolución atómica. Para crear este modelo, los autores han realizado a buen seguro un concienzudo estudio de las estructuras de proteínas determinadas experimentalmente. Al tratarse de un tipo de giro que no aparece con frecuencia en proteínas, el modelo no está diseñado para reproducirlo. Por otro lado, también por su poca representatividad, podemos pensar que esto no supone un problema importante.

La ventaja más significativa de los modelos simplificados frente a los más detallados es que requieren menos recursos computacionales. Para cada uno de los modelos hemos medido el tiempo que le necesita el programa para realizar una serie de operaciones. Concretamente, hemos medido el tiempo que tarda cada uno de los programas en decodificar una población de 1000 cromosomas, generar las correspondientes conformaciones y calcular sus energías. Para el modelo de Chen, el tiempo que se invierte es 0.8 veces el tiempo que tarda el programa del modelo de Irbäck. El modelo de Kolinski tarda 1.2 veces lo que tarda el modelo de Irbäck. Por tanto, el modelo más barato computacionalmente es el de Chen, seguido del de Irbäck. La diferencia entre ellos se debe al número de centros por residuo, mayor para el de Irbäck, y al cálculo de la energía, que en el caso de Chen no tiene componente orientacional. El modelo de Kolinski resulta más caro debido al gran número de vectores que han de ser calculados y, sobre todo, a las restricciones

que preceden al cómputo de la energía. Sin embargo, estos resultados para el tiempo de cálculo son sólo orientativos. No se corresponden con lo que se obtendría si se realizasen simulaciones de tipo Monte Carlo del plegamiento, con más contribuciones para la energía y un muestreo conformacional en que se considerasen otros grados de libertad. Por ejemplo, costaría más realizar movimientos de Monte Carlo con el modelo de Irbäck que con los otros dos. Los grados de libertad del primero son los ángulos de torsión reales de la cadena, mientras que para el de Chen y el de Kolinski son ángulos virtuales. Por otro lado, reasignar las posiciones de los átomos PB del modelo de Kolinski es trivial, mientras que con los modelos de Chen e Irbäck requeriría más tiempo. Por ejemplo, en el caso del modelo de Chen, en cada paso habría que recalcular las posiciones de los átomos virtuales H' y O', algo bastante caro computacionalmente. Además, el número total de cálculos de energía sería muy superior en simulaiones más detalladas que con un muestreo conformacional simplificado como el nuestro.

Este estudio nos ha permitido conocer muchas propiedades de los modelos simplificados para el enlace de hidrógeno. Estos modelos son una herramienta muy importante para el estudio por simulación molecular del plegamiento de proteínas. Sin embargo, ninguna de sus simplificaciones —en la representación de la proteína o en las funciones energéticas— es gratuita, por lo que debe prestarse mucha atención a su influencia. Nuestros resultados dirigen la atención hacia la manera en que estas simplificaciones de los modelos pueden resultar problemáticas. Aun así, podemos decir que los modelos de grano grueso evaluados funcionan muy bien reproduciendo los efectos de las interacciones de tipo enlace de hidrógeno en proteínas.

Capítulo 6

Evaluación conjunta de los potenciales hidrófobo y de enlace de hidrógeno

En los Capítulos 4 y 5 hemos estudiado una serie de potenciales para la interacción hidrófoba y para el enlace de hidrógeno del esqueleto de proteínas, respectivamente. Hemos realizado estos estudios independientemente, es decir, evaluando de manera aislada los potenciales para los distintos tipos de interacción. En este Capítulo estudiamos conjuntamente los mejores potenciales de las evaluaciones precedentes. Como hemos visto en el Capítulo 4, el potencial hidrófobo que mejores resultados ofrece de los tres que hemos probado es el DFIRE-SCM de Zhou *et al*[123]. Entre los potenciales de enlace de hidrógeno que hemos analizado[102–104], establecer cuál es más apropiado para nuestro estudio es más complicado, dado que adquieren mayor importancia consideraciones como la resolución de los distintos modelos. Por eso dedicamos una sección de este Capítulo a seleccionar el potencial de enlace de hidrógeno que combinamos con el potencial hidrófobo.

Para poder calcular la energía de una conformación de la proteína considerando más de una contribución, uno de los aspectos que hay que tratar es el peso relativo que

se asigna a cada tipo de interacción. En los campos de fuerza de mecánica molecular este ajuste se lleva a cabo utilizando una gran cantidad de información sobre la magnitud de las interacciones, tanto experimental como obtenida a partir de cálculos *ab initio*[39–42]. En estudios más sencillos en los que también se consideran varios tipos de interacción, el ajuste de las distintas contribuciones se realiza a menudo a partir de una serie de cálculos preliminares sobre un conjunto pequeño de proteínas[102,103]. Como también explicamos más adelante en este Capítulo, vamos a usar como referencia los valores experimentales para los distintos tipos de interacción y estudios realizados con modelos sencillos para los mismos.

Finalmente, llevamos a cabo una nueva serie de experimentos de minimización de la energía para un conjunto de proteínas. Entre ellas, consideramos proteínas todo α y todo β, con las que comprobamos si los potenciales retienen su buen comportamiento al añadir otra contribución a la energía. Además, realizamos minimizaciones con proteínas de tipo $(\alpha + \beta)$, donde las dos contribuciones a la energía tienen una importancia comparable.

6.1. Selección de un potencial de enlace de hidrógeno

En el Capítulo 5 hemos estudiado tres potenciales para la interacción de enlace de hidrógeno, haciendo especial hincapié en el nivel de resolución en la representación del esqueleto de la proteína. El primero de los potenciales que hemos evaluado es el de Irbäck[102], de resolución atómica para el esqueleto. Hemos utilizado este potencial como referencia con la que comparar otros dos mucho más simplificados, el de Chen[104] y el de Kolinski[103]. Por su baja resolución, estos dos últimos modelos ofrecen resultados muy favorables en nuestra evaluación y pueden ser apropiados para realizar simulaciones del plegamiento. Sin embargo, con ambos modelos hemos observado una serie de posibles

problemas derivados de su simplicidad. Para hacer una estimación del comportamiento que podemos esperar de ellos en un estudio con otra contribución a la energía, hemos realizado una serie de experimentos de minimización sumando la contribución del potencial de Chen o el de Kolinski y el potencial DFIRE-SCM. Como es habitual, hemos llevado a cabo estos experimentos con nuestro método evolutivo.

En la Figura 6.1 mostramos resultados de algunos de estos experimentos para la combinación del potencial DFIRE-SCM y el potencial de Chen (a la que a partir de ahora llamamos DFIRE-Chen). En cada panel representamos energía frente a $RMSD$ respecto a la nativa de las mejores conformaciones a lo largo de la optimización. Los dos primeros paneles corresponden a una proteína todo α, cuyo código PDB es 1i6z, y otra todo β, 1tk7, ambas empleadas anteriormente (ver Capítulos 4 y 5). En ambos casos, la combinación de potenciales DFIRE-Chen define el mínimo energético en una conformación con $RMSD$ frente a la nativa próximo a cero. Por tanto, el test mínimo de que se conserven los buenos resultados de las contribuciones aisladas se verifica para el potencial DFIRE-Chen.

El tercero de los paneles de la Figura 6.1 corresponde a una proteína de tipo (α+β), 1agt, en cuya conformación nativa aparece una lámina β empaquetada contra una hélice α. En este caso, el potencial DFIRE-Chen no es capaz de definir el mínimo en la conformación nativa. Como podemos ver en la Figura, la conformación de menor energía aparece con $RMSD$ próximo a 8 Å. En la Figura 6.2 mostramos los mapas de enlace de hidrógeno según el modelo de Chen para la conformación nativa (a) y para la optimizada (b). Dividimos cada uno de los mapas en sectores conforme al número de fragmentos con estructura secundaria bien definida que encontramos en la conformación nativa de la proteína. Si seguimos la secuencia de la proteína, el orden de los distintos elementos de estructura secundaria es β1-α-β2-β3. En el mapa correspondiente a la conformación nativa vemos que el modelo detecta varias repulsiones, debido al volumen

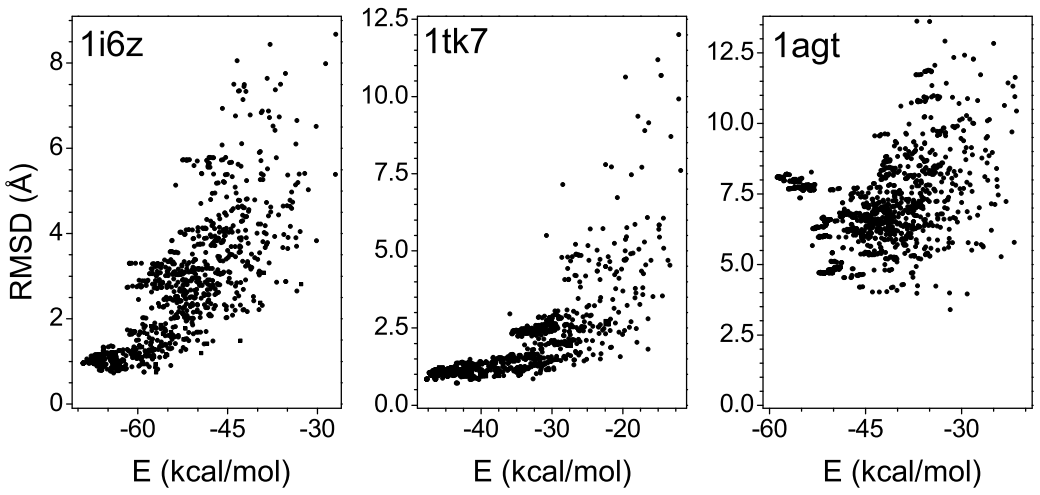


Figura 6.1: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas 1i6z, 1tk7 y 1acw, en la minimización con la combinación de potenciales DFIRE-SCM y Chen.

excluido entre carbonos-α del modelo. Por otra parte, encontramos los patrones de contactos propios de los distintos tipos de estructura secundaria, tal y como los hemos descrito en el Capítulo 5. Los patrones de contactos de la hélice aparecen entre residuos del fragmento 2, y los de la lámina, entre las hebras $\beta 1$ y $\beta 3$, y entre las hebras $\beta 2$ y $\beta 3$. Por tanto, el modelo de Chen captura los enlaces de hidrógeno nativos correctamente, aunque aparecen, como también veíamos en el Capítulo 5, contactos hélice-hebra más débiles. Si miramos ahora al mapa correspondiente a la conformación optimizada con el potencial DFIRE-Chen, observamos que prácticamente se ha perdido el patrón de la lámina. En esta conformación, las interacciones más importantes para la contribución de enlace de hidrógeno de Chen son las que se forman entre la hélice α y las hebras $\beta 1$ y $\beta 2$. Además, aparecen interacciones no nativas entre las hebras $\beta 1$ y $\beta 3$. En la definición de este mínimo energético con el potencial DFIRE-Chen, la contribución de Chen está contribuyendo al colapso entre la hélice y las hebras, en lugar de permitir que se recupere la lámina. Este colapso, al contrario del que favorece la contribución

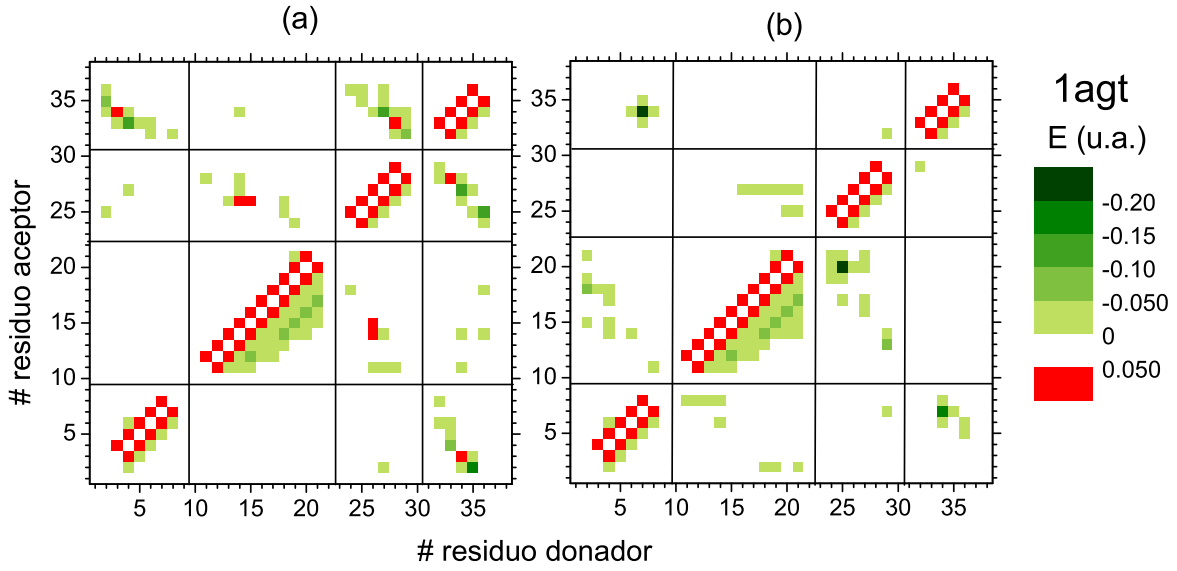


Figura 6.2: Mapas de energía de dos conformaciones de la proteína 1agt calculados con el potencial de Chen: (a) conformación nativa y (b) conformación de mínima energía obtenida en la optimización con el potencial DFIRE-Chen. Los valores de energía se expresan en unidades del modelo.

del potencial DFIRE-SCM, es inespecífico con la secuencia, y se debe a la falta de una dependencia orientacional en el modelo, ya descrita en el Capítulo 5.

Los resultados preliminares obtenidos con el potencial DFIRE-Chen, entre los que se incluyen los de otras estructuras que no comentamos aquí, sugieren que la definición de la superficie de energía pueda verse fuertemente afectada por la problemática del potencial de Chen. En adelante nos centramos en la combinación de potenciales DFIRE-SCM y Kolisnki (a partir de ahora, DFIRE-Kolinski). Con esta combinación de potenciales hemos obtenido mejores resultados preliminares que con la combinación DFIRE-Chen. Por eso, lo hemos utilizado en una serie de experimentos de minimización sobre un conjunto más amplio de proteínas. Los resultados de los experimentos, en los que se incluyen los obtenidos para las proteínas 1i6z, 1k43 y 1agt, se incluyen más adelante en este mismo Capítulo.

6.2. Parametrización del potencial DFIRE-Kolinski

Para estudiar conjuntamente los potenciales que hemos seleccionado, el potencial DFIRE-SCM y el término de enlace de hidrógeno de Kolinski, debemos asignarle a cada uno de ellos un peso en la energía global de la proteína. Como hemos explicado en los correspondientes capítulos, estos dos potenciales tienen origen muy dispar. El potencial DFIRE-SCM se deduce a partir de un tratamiento estadístico de los datos experimentales de estructuras de proteínas[123]. Una serie de aproximaciones, entre las que se incluye la comparación con valores experimentales de estabilidad[123], permiten que los términos energéticos entre pares se expresen con unidades. El potencial de enlace de hidrógeno de Kolinski también procede de las estructuras de proteínas, pero sólo en la medida en que permite recuperar estructuras como hélices α y láminas β a partir de una serie de correlaciones geométricas[103]. Como hemos visto en el Capítulo 5, con este modelo la energía de enlace de hidrógeno se obtiene en unidades del propio modelo, no comparables directamente con las del potencial DFIRE-SCM. Para poder calcular la energía global de un conformación para una proteína con la combinación DFIRE-Kolinski debemos dar un peso relativo a cada una de las dos contribuciones. Así, para darle unidades a la energía de enlace de hidrógeno en el modelo multiplicamos su valor por un factor ω_{hb}. Este factor es el cociente entre una estimación de la estabilización por la formación de cada enlace de hidrógeno (entre -3 y -10 kcal/mol)[24,25] y la energía mínima que puede alcanzar un enlace de hidrógeno en unidades del modelo. Así, podemos obtener un valor para la contribución de enlace de hidrógeno que puede sumarse a la componente hidrófoba.

A pesar de lo aproximada que es la parametrización que introducimos, tenemos motivos para pensar que los resultados puedan no depender dramáticamente de los pesos para las distintas contribuciones, al menos en experimentos de minimización como los que realizamos. Como hemos comentado con anterioridad, las proteínas son sistemas

mínimamente frustrados[22]. Esto significa que las diversas interacciones que estabilizan las proteínas no están en conflicto unas con otras. Por el contrario, actúan sinérgicamente para definir el mínimo energético en la forma activa de la proteína de una manera robusta, es decir, insensible a mutaciones puntuales en la secuencia. Por tanto, podemos pensar que mientras los distintos tipos de interacción que consideremos estén representados significativamente, la capacidad de definir el mínimo en la conformación nativa puede no ser muy sensible a los pesos de cada una de las contribuciones. Para comprobar esta robustez del potencial ante pequeños cambios en la parametrización hemos realizado unas pruebas con nuestro método de optimización con dos valores diferentes para la energía del enlace de hidrógeno, ambas dentro del intervalo mencionado. En los resultados de estos cálculos hemos observado que, en efecto, ante estos cambios no hay variaciones importantes en la definición del mínimo energético con el potencial DFIRE-Kolinski. Para realizar los cálculos cuyos resultados se presentan en las siguientes secciones hemos utilizado el valor $\omega_{hb} = 5$ kcal/mol.

6.3. Minimización de la energía con el potencial DFIRE-Kolinski

Como hemos venido haciendo a lo largo de esta Tesis, para estudiar el potencial DFIRE-Kolinski hemos llevado a cabo una serie de experimentos de minimización con nuestro algoritmo evolutivo. En este caso, la función de mérito que el método minimiza es la suma de las dos contribuciones a la energía, calculada de acuerdo con lo comentado en la sección anterior.

En los capítulos anteriores, para la evaluación de los potenciales hemos utilizado proteínas cuya conformación nativa es de un determinado tipo estructural. Hemos escogido un tipo de estructura u otro en función de la interacción que se estudiara. En

concreto, en el caso de los potenciales hidrófobos hemos utilizado proteínas todo α (ver Capítulo 4), y en el caso de los potenciales de enlace de hidrógeno, proteínas todo β (Capítulo 5). Utilizando nuestro método de ensamblaje de fragmentos rígidos, un potencial bien diseñado para uno de estos tipos de interacción debe ser capaz de asignar la mínima energía a la conformación nativa para el tipo de estructura correspondiente. En esta parte de nuestro estudio, hemos querido comprobar si con la combinación de potenciales DFIRE-Kolinski se conservan los buenos resultados obtenidos independientemente para las dos contribuciones.

Así, en primer lugar hemos llevado a cabo la minimización de la energía para dos proteínas todo α y dos todo β ya utilizadas en los Capítulos 4 y 5, respectivamente. Mostramos estas cuatro proteínas en la Figura 6.3. Cuando calculamos la energía para proteínas de estos dos tipos con el potencial DFIRE-Kolinski esperamos que una de las contribuciones sea mayoritaria en la energía global de la proteína. Lógicamente, esta contribución será la hidrófoba en el caso de las proteínas todo α, y la de enlace de hidrógeno cuando ensamblamos hebras β. Para completar el análisis hemos seleccionado además una serie de proteínas $(\alpha + \beta)$, en cuya conformación nativa aparecen representados los dos tipos principales de estructura secundaria. En la Figura 6.4 mostramos las proteínas de este tipo que vamos a estudiar. Con este conjunto de proteínas $(\alpha + \beta)$ podemos conocer el funcionamiento de los potenciales cuando las dos contribuciones son aproximadamente igual de relevantes.

Como en los capítulos precedentes, para cada proteína hemos tomado la secuencia y las coordenadas de su conformación nativa del Protein Data Bank[56,57]. Para utilizarlas en nuestro método hemos dividido cada una de las proteínas en tantos fragmentos como hélices α o hebras-β la formen, de tal modo que quede un solo enlace peptídico entre fragmentos. Por este motivo, una vez más utilizamos la codificación interna en el algoritmo evolutivo (ver Capítulos 2 y 3). Para todas las proteínas hemos realiza-

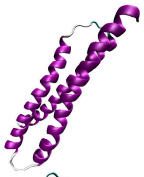
Descripción	Código PDB	nº fragmentos	nº residuos en el modelo	
1. Proteínas todo α				
Chaperona reguladora de la familia BAG	**1i6z**	3	116	
Dominio FAT de la quinasa de adhesión focal	**1ktm**	4	128	
2. Proteínas todo β				
Péptido MBH12 de 14 residuos RG-KWTY-NG-ITYE-GR	**1k43**	2	14	
Dominios WW3-4 del supresor de deltex	**1tk7**	3	25	

Figura 6.3: Conjunto de proteínas todo α y todo β para el estudio del potencial DFIRE-Kolinski, con el número de fragmentos en el modelo, el número de residuos y una representación de su estructura tridimensional.

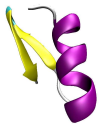
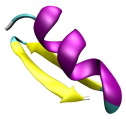
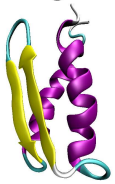
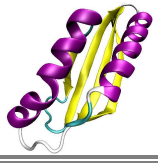
Descripción	Código PDB	nº fragmentos	nº residuos en el modelo	
3. Proteínas (α + β)				
Miniproteína que reproduce el núcleo de CD4	**1d5q**	3	27	
Péptido natural de escorpión P01	**1acw**	3	29	
Defensina de mejillón MGD1	**1fjn**	3	39	
Agitoxina 2 inhibidora de canales de potasio	**1agt**	4	38	
Butantoxina	**1c56**	4	40	
Proteína W del bacteriófago λ	**1hyw**	4	58	
Dominio de unión de la proteína G de estreptococo	**2gb1**	5	56	
Proteína TT1725 de Thermus thermophilus	**1j27**	6	98	

Figura 6.4: Conjunto de proteínas (α+β) para el estudio conjunto del potencial DFIRE-Kolinski, con el número de fragmentos en el modelo, el número de residuos y una representación de su estructura tridimensional.

do minimizaciones de la energía sin modificar los parámetros del algoritmo evolutivo. Como en partes anteriores de nuestro estudio, adaptamos el número de minimizaciones independientes de cada ejecución del algoritmo. Así, para las proteínas que dividimos en dos o tres fragmentos, cada ciclo de optimización está formado por cinco minimizaciones independientes; son diez para las proteínas que tienen cuatro o cinco fragmentos, y quince para las proteínas de mayor tamaño. Cada una de las minimizaciones la hemos repetido cinco veces con distintos conjuntos de números semilla.

6.4. Resultados de la minimización con el potencial DFIRE-Kolinski

6.4.1. Resultados de la minimización para proteínas todo α

Como hemos comentado, para nuestro estudio del potencial DFIRE-Kolinski consideramos en primer lugar dos proteínas todo α, con códigos PDB 1i6z y 1ktm. Al igual que en el Capítulo 4, dividimos su estructura nativa en fragmentos rígidos de nuestro modelo de tal modo que cada uno contiene una de las hélices α de la proteína. La energía se calcula con el potencial DFIRE-Kolinski únicamente entre estos fragmentos. Por eso en este caso la contribución mayoritaria es la del colapso hidrófobo entre hélices, mientras que la aportación a la estabilidad del enlace de hidrógeno es marginal. En la Tabla 6.1 recogemos valores de energía para las conformaciones nativas de 1i6z y 1ktm, que corroboran esta apreciación. En ambos casos la energía de enlace de hidrógeno de la conformación nativa, calculada con el potencial de Kolinski (E_{Nat}^{Kol}), es cero o casi cero.

En la minimización de la energía, tanto para 1i6z como para 1ktm se alcanzan conformaciones muy parecidas a la nativa, como indican los valores de $RMSD$ de la Tabla 6.1, con energía levemente inferior a la de la nativa. Como esperábamos, también

$PDBid$	n^o $frag$	E_{Nat}	E_{Nat}^{DFIRE}	E_{Nat}^{Kol}	E_{Min}	E_{Min}^{DFIRE}	E_{Min}^{Kol}	$RMSD$
1i6z	3	-52.8	-52.6	-0.2	-65.6	-65.6	0.0	0.8
1ktm	4	-89.0	-89.0	0.0	-108.4	-104.4	-4.0	0.7

Tabla 6.1: Resultados de la optimización con el potencial DFIRE-Kolinski para proteínas todo α: energía de la conformación nativa (E_{Nat}) y del mínimo energético (E_{Min}), ambas en kcal/mol, y valor de $RMSD$ en angstroms del mínimo con respecto a la nativa. Los superíndices Kol y $DFIRE$ indican la contribución a la que corresponde la energía.

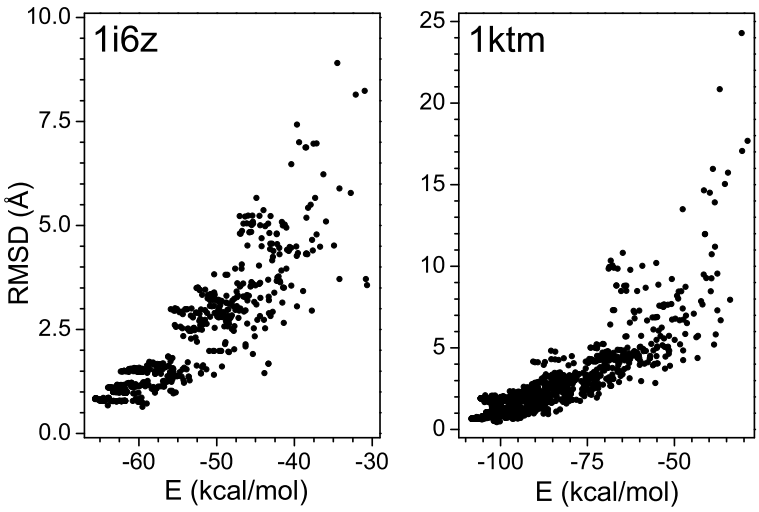


Figura 6.5: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas todo α 1i6z y 1ktm en la minimización con el potencial DFIRE-Kolinski.

en las conformaciones de mínima energía de las dos proteínas la contribución de enlace de hidrógeno es muy pequeña. En la Figura 6.5 mostramos una representación de energía frente a $RMSD$ frente a la estructura nativa de las conformaciones de menor energía que van encontrándose a lo largo de la optimización para las dos proteínas. Tanto para 1i6z como para 1ktm, la definición de la superficie permite que el algoritmo encuentre el mínimo energético con facilidad. Por tanto, se mantienen las buenas propiedades del potencial DFIRE-SCM para definir el mínimo energético en la conformación nativa con proteínas todo α que explicamos en el Capítulo 4.

6.4.2. Resultados de la minimización para proteínas todo β

Los resultados que hemos obtenido con las dos proteínas todo β que hemos seleccionado, 1k43 y 1tk7, son semejantes a los que acabamos de describir para las proteínas todo α. Como explicamos en el Capítulo 5, los fragmentos que consideramos en nuestro modelo son las hebras β, que forman una lámina estabilizada principalmente por enlaces de hidrógeno. Por tanto, en este caso la contribución minoritaria ha de ser la hidrófoba. Esto lo comprobamos en los valores de energía de la contribución del potencial DFIRE-SCM (E_{Nat}^{DFIRE}) de la Tabla 6.2. Aun así, la aportación a la energía global de la contribución minoritaria es más importante aquí que en el caso de las proteínas todo α. Esto es debido a que en las láminas β las cadenas laterales de los residuos interaccionan entre sí tanto por encima como por debajo del plano de la lámina.

También en el caso de las proteínas todo β, al llevar a cabo la minimización de la energía, alcanzamos conformaciones muy parecidas a las correspondientes nativas, como indican los valores de $RMSD$ de la Tabla 6.2, próximos a cero. En este caso, el cambio en la energía con respecto a la nativa es más importante que en el caso de las proteínas todo α, llegando los valores de E_{Min}^{Kol} casi a doblar los de E_{Nat}^{Kol}. Esto lo observábamos también en las minimizaciones sólo con el potencial de Kolinski que hemos descrito en el Capítulo 5. Como hemos comentado en el análisis de este potencial, las restricciones que impone para definir enlaces de hidrógeno son muy estrictas. Por este motivo, en ocasiones, no es capaz de detectar interacciones entre residuos que sí forman enlaces de

$PDBid$	n^o $frag$	E_{Nat}	E_{Nat}^{DFIRE}	E_{Nat}^{Kol}	E_{Min}	E_{Min}^{DFIRE}	E_{Min}^{Kol}	$RMSD$
1k43	2	-12.7	-1.8	-11.0	-21.6	-2.0	-19.7	0.6
1tk7	3	-33.4^3	-4.4	-28.5	-49.5	-3.6	-45.9	0.8

Tabla 6.2: Resultados de la optimización con el potencial DFIRE-Kolinski: energía de la conformación nativa y del mínimo energético, ambas en kcal/mol, y valor de $RMSD$ en angstroms del mínimo con respecto a la nativa. El superíndice numérico en los valores de energía indica el número de repulsiones en la conformación nativa.

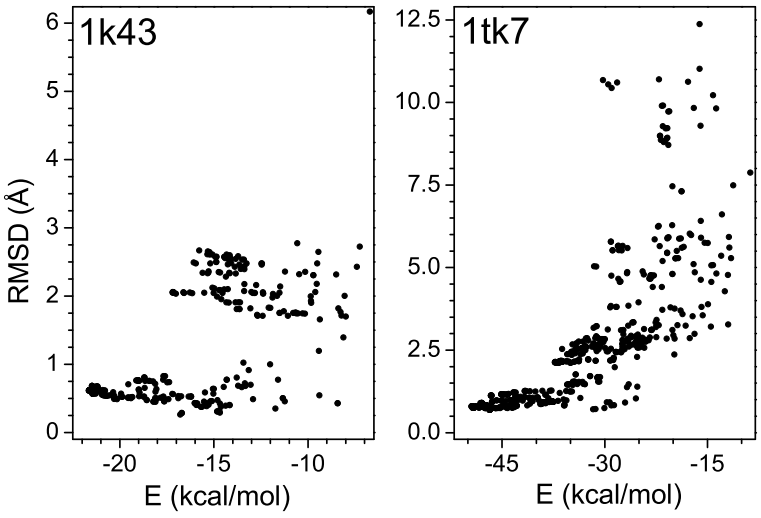


Figura 6.6: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas todo β 1k43 y 1tk7, en la minimización con el potencial DFIRE-Kolinski.

hidrógeno en la conformación nativa. Al llevar a cabo la minimización se encuentran conformaciones muy próximas a la nativa en las que sí se capturan estas interacciones, y que por tanto tienen un valor de energía muy inferior para esta contribución. Esta mejora en el término de enlace de hidrógeno se produce, a veces, a costa de un leve empeoramiento en el valor de la contribución hidrófoba.

En la Figura 6.6 mostramos representaciones de energía frente a $RMSD$ con respecto a la nativa de las conformaciones de menor energía que se encuentran en la optimización según van transcurriendo las generaciones para las proteínas 1k43 y 1tk7. El potencial es capaz de definir una superficie de energía en la que el algoritmo encuentra fácilmente el mínimo absoluto en las vecindades de la conformación nativa. Por tanto, al incorporar una nueva contribución a la energía se conserva el buen comportamiento del potencial de enlace de hidrógeno de Kolinski.

6.4.3. Minimización de la energía con proteínas ($\alpha + \beta$)

Además de las pruebas con proteínas todo α y todo β, consideramos una serie de proteínas de tipo (α+β) con distinto número de fragmentos en nuestro modelo (ver Figura 6.4). En la estructura nativa de las proteínas que hemos seleccionado encontramos distinto número de hélices α, entre una y dos, y distinto número de hebras β, entre dos y cuatro formando una sola lámina. Las hélices se encuentran empaquetadas contra la superficie de la lámina, por lo que estas proteínas suponen un test óptimo para poner a prueba el potencial DFIRE-Kolinski.

En la Tabla 6.3 mostramos los resultados obtenidos para este conjunto de proteínas. Si atendemos a la energía de las estructuras nativas observamos que las dos contribuciones son significativas en la energía global. Aun así, en la mayoría de los casos la contribución mayoritaria es la del potencial hidrófobo (E_{Nat}^{DFIRE}). Esta tendencia se observa para todas las proteínas que estudiamos excepto 2gb1, en cuya estructura encontramos hasta cuatro hebras β. Para esta proteína el término de enlace de hidrógeno (E_{Nat}^{Kol}) tiene un valor superior, debido a que el modelo de Kolinski captura muy adecuadamente los enlaces de hidrógeno nativos entre sus hebras β, como hemos visto en

$PDBid$	n^o $frag$	E_{Nat}	E_{Nat}^{DFIRE}	E_{Nat}^{Kol}	E_{Min}	E_{Min}^{DFIRE}	E_{Min}^{Kol}	$RMSD$
1d5q	3	-33.8	-20.4	-13.4	-40.5	-16.9	-23.9	0.6
1acw	3	-22.0^{6}	-14.4	-7.6	-44.4	-19.1	-25.3	1.1
1fjn	3	-27.1^{4}	-22.9	-4.2	-34.4	-21.9	-12.5	3.6
1agt	4	-49.3^{2}	-33.2	-16.1	-63.1	-29.6	-33.7	1.0
1c56	4	-39.4^{5}	-28.4	-11.0	-54.5	-31.0	-23.5	2.1
1hyw	4	-27.9^{5}	-21.2	-6.7	-57.6	-30.7	-26.9	3.8
2gb1	5	-74.8^{2}	-31.8	-43.0	-99.7	-32.9	-66.8	1.5
1j27	6	-125.6	-78.4	-47.1	-135.7	-56.4	-79.3	2.3

Tabla 6.3: Resultados de la optimización con el potencial DFIRE-Kolinski para proteínas ($\alpha + \beta$): energía de la conformación nativa (E_{Nat}) y el mínimo energético (E_{Min}), ambas en kcal/mol, y valor de $RMSD$ en angstroms del mínimo con respecto a la nativa. El superíndice numérico en los valores de energía indica el número de repulsiones en la conformación nativa.

la Sección 5.2.3.

Para la mayoría de las proteínas seleccionadas, con el potencial DFIRE-Kolinski localizamos repulsiones de volumen excluido en la conformación nativa (indicadas como un superíndice en los valores de E_{Nat} de la Tabla 6.3). La mayoría de estas repulsiones se encuentran entre uno de los átomos virtuales PB del modelo CABS[103] (ver Capítulo 5) y otro átomo del modelo de la proteína, bien un carbono-α o bien otro centro PB. En ocasiones, estas repulsiones se detectan entre dos carbonos-α, como sucede para las proteínas 1acw, 1c56 y 1hyw. El gran número de repulsiones en las estructuras nativas se debe, por una parte, a las estrictas restricciones del modelo de Kolinski. Por otra parte, en algunas de estas proteínas hay puentes disulfuro entre residuos de cisteína que en nuestro modelo forman parte de distintos fragmentos. Los puentes disulfuro son muy frecuentes en proteínas de secreción de pequeño tamaño, como algunas de las que utilizamos en esta parte de nuestro estudio[1]. Estas interacciones de tipo covalente propician una excepcional proximidad entre las regiones en que se encuentran las cisteínas implicadas. En la Figura 6.7 mostramos una representación de la proteína 1acw, en la que destacamos los tres puentes disulfuro que acercan las regiones con distinto tipo de estructura secundaria. Debido a esta cercanía, los potenciales detectan repulsiones de volumen excluido en su conformación nativa.

Para todas estas proteínas hemos llevado a cabo minimizaciones energéticas con el potencial DFIRE-Kolinski. Si atendemos a los valores de $RMSD$ de la conformación optimizada respecto a la nativa para las proteínas que hemos estudiado (ver Tabla 6.3), podemos decir que los resultados de la minimización son muy buenos. Para algunas de estas proteínas se ha alcanzado un mínimo con $RMSD$ próximo a 1 Å (1d5q, 1acw, 1agt y 2gb1), y en algunos de los casos en que los que el valor de $RMSD$ es superior a 2 Å (1c56 y 1j27) el mínimo también es muy parecido a la estructura nativa. Esto supone un indudable éxito para el potencial DFIRE-Kolinski, en el que el hemos uti-

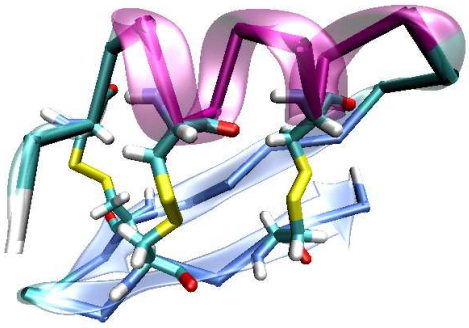


Figura 6.7: Representación de la conformación nativa de la proteína 1acw. Destacamos en amarillo los tres puentes disulfuro que se forman entre residuos de cisteína situados en las hebras β y la hélice α.

lizado una parametrización sumamente sencilla para sumar las dos contribuciones a la energía. Además, como ya hemos comentado, los buenos resultados no parecen depender fuertemente del peso exacto de cada una de las contribuciones mientras ambas estén suficientemente representadas en la energía global.

A pesar de la semejanza entre la conformación nativa y la minimizada, para las ditintas proteínas observamos importantes diferencias en los valores de energía (ver Tabla 6.3). Estos cambios se deben principalmente a la contribución de enlace de hidrógeno de Kolinski, que en el mínimo energético es mucho mayor en valor absoluto. Como hemos visto para las proteínas todo β, el potencial de Kolinski no siempre captura bien los enlaces de hidrógeno nativos entre hebras. Sin embargo, en una disposición de los fragmentos muy parecida, en la que además se evitan las repulsiones de volumen excluido, el potencial sí es capaz de identificar los enlaces de hidrógeno que estabilizan la lámina nativa. Como hemos visto para las proteínas todo β, esta mejora en el valor de la energía de enlace de hidrógeno se consigue a menudo a costa de un leve empeoramiento de la componente hidrófoba. Así, los valores de esta contribución en la estructura nativa (E_{Nat}^{DFIRE}) para las proteínas 1d5q, 1fjn, 1agt y 1j27 son menores que en la minimizada (E_{Min}^{DFIRE}). En todo caso, las fluctuaciones entre los valores de energía de la conformación

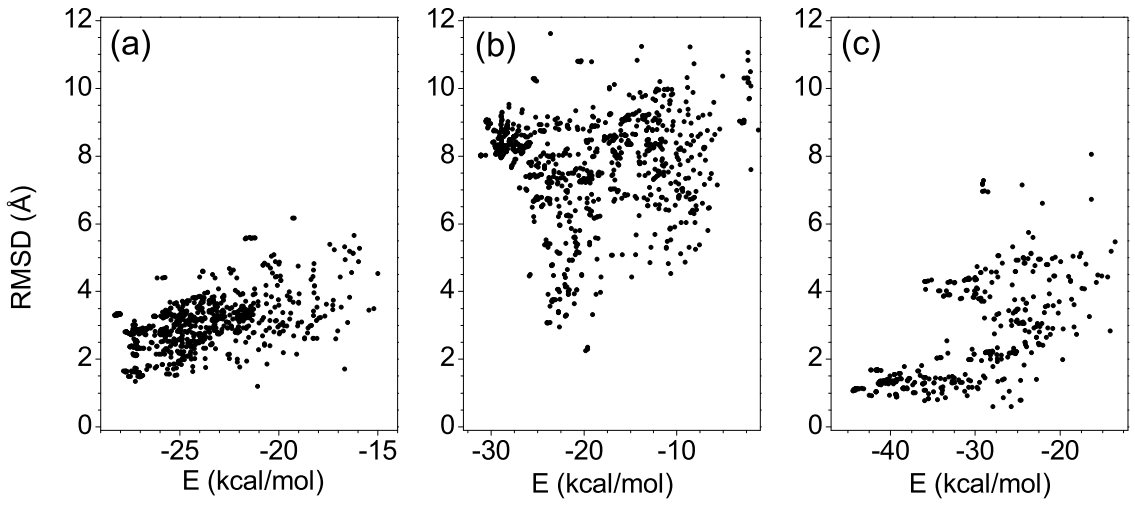


Figura 6.8: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de la proteína 1acw en minimizaciones con distintas funciones de mérito: (a) potencial DFIRE-SCM, (b) potencial de Kolinski, y (c) potencial DFIRE-Kolinski.

nativa y la minimizada no impiden que en la mayoría de casos el mínimo se defina correctamente en la conformación nativa. Esto se debe a que las dos contribuciones actúan sinérgicamente para definir el mínimo en esta conformación.

Hemos querido comprobar con mayor detalle si existe este efecto cooperativo entre los dos potenciales con una serie de experimentos realizados con algunas de las proteínas $(\alpha + \beta)$ que estamos considerando. En la Figura 6.8 mostramos resultados para 1acw, una proteína de tres fragmentos en nuestro modelo. Las distintas minimizaciones difieren en la función de mérito utilizada: solo el potencial DFIRE-SCM (a), solo el potencial de Kolinski (b), y la suma las dos contribuciones (potencial DFIRE-Kolinski) (c). Con el potencial DFIRE-SCM se alcanza un conjunto de conformaciones de energía muy parecida, con $RMSD$ entre 1.5 y 4 Å frente a la nativa (Figura 6.8 (a)). Estas conformaciones corresponden a empaquetamientos de los tres fragmentos que maximizan los contactos, en general con más espacio entre las hebras β. Aunque estas conformaciones

no son muy diferentes de la nativa, la definición del mínimo presenta una importante degeneración. En el caso de la minimización únicamente con el potencial de Kolinski, el mínimo se define en una conformación con $RMSD$ próximo a 8 Å (Figura 6.8 (b)). En esta conformación, la horquilla β está bien formada pero la hélice prácticamente no interacciona con ella. El único caso en el que la superficie de energía dirige el muestreo hacia un mínimo bien definido en la conformación nativa es cuando usamos el potencial DFIRE-Kolinski (Figura 6.8 (c)). A partir de resultados como estos, semejantes a los obtenidos para otras proteínas que no mostramos aquí, se comprueba que las dos contribuciones actúan de manera coordinada para definir correctamente el mínimo.

Esta capacidad de definir una superficie de energía que dirija la minimización hacia una conformación parecida a la nativa la observamos para muchas de las proteínas $(\alpha + \beta)$ que estamos considerando. En la Figura 6.9 mostramos representaciones de energía frente a $RMSD$ de las conformaciones de menor energía que se obtienen a lo largo de las generaciones para todas ellas. En el caso de las proteínas 1d5q, 1acw, 1agt, 2gb1 y 1j27, que tienen entre tres y seis fragmentos peptídicos en nuestro modelo, la superficie permite que se alcance el mínimo nativo con relativa facilidad. En los casos restantes, 1fjn, 1c56 y 1hyw, o bien la superficie de energía no está tan bien definida o bien el mínimo no se parece tanto a la conformación nativa, por razones que explicaremos más adelante en esta misma sección.

Para dos de las proteínas de tres fragmentos hemos obtenido resultados muy buenos. Se trata de 1acw, de la que ya hemos hablado, y 1d5q. En el caso de 1d5q aparece un segundo mínimo local, más pobremente definido que el nativo, en el que las hebras adquieren una disposición levemente diferente con respecto a la hélice. Sin embargo, finalmente se alcanza sin dificultad el mínimo nativo gracias a la mejor definición de este pozo en la superficie de energía (ver Figura 6.9). Para la tercera proteína de tres fragmentos, 1fjn, la situación es la contaria a la de 1d5q. En este caso aparecen dos

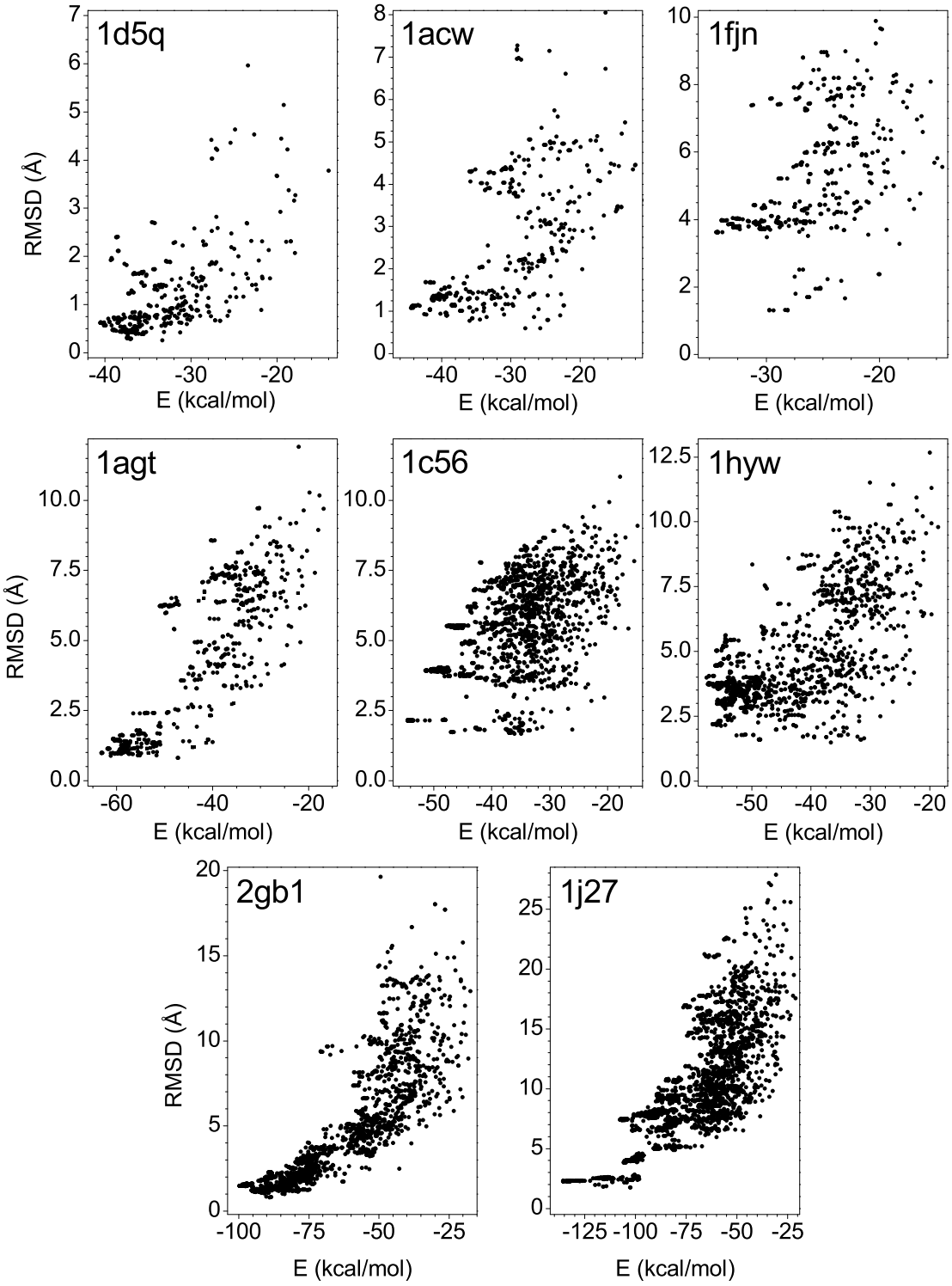


Figura 6.9: Representación de energía frente a $RMSD$ con respecto a la nativa en angstroms de las mejores conformaciones de las proteínas $(\alpha + \beta)$ consideradas, en la minimización con la combinación de potenciales DFIRE-SCM y Kolinski.

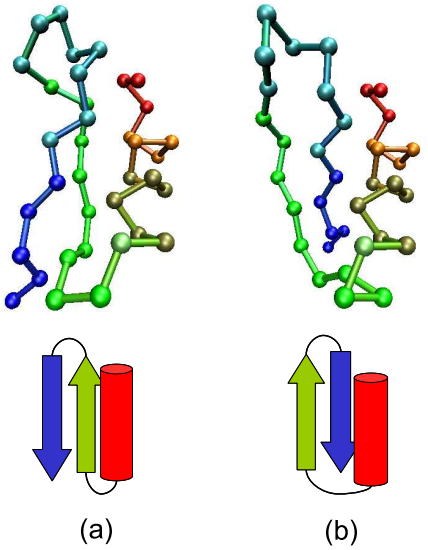


Figura 6.10: Representación simplificada y diagrama topológico para dos conformaciones de 1fjn: (a) conformación nativa, y (b) mínimo energético con $RMSD = 3.6$ Å.

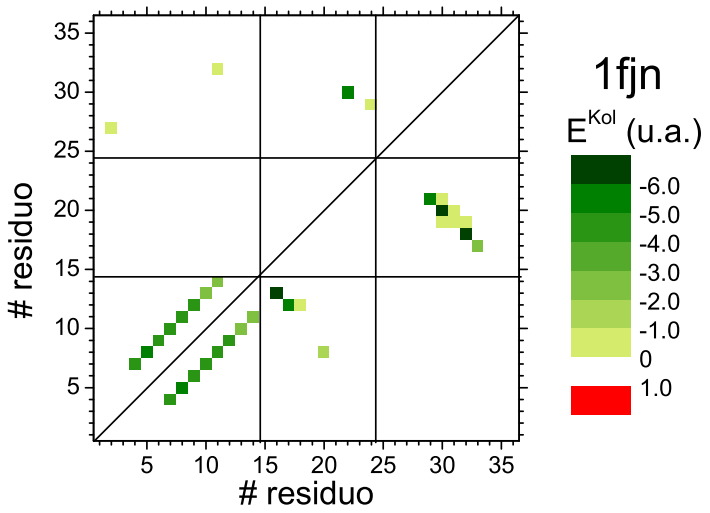


Figura 6.11: Mapa de energías de 1fjn calculado con el potencial de Kolinski para las conformaciones nativa (triángulo superior) y optimizada con el potencial DFIRE-Kolinski (triángulo inferior). La energía se expresa en unidades del modelo.

mínimos, entre los cuales el peor definido es el nativo. La minimización alcanza una conformación con $RMSD \simeq 4$ Å con respecto a la nativa. Mostramos una representación simplificada de ambas conformaciones de 1fjn en la Figura 6.10. La conformación optimizada (b) corresponde a una disposición en que se permuta la posición relativa de las hebras con respecto a su disposición en la nativa (a). La mejor definición de este mínimo no nativo se debe a un conjunto de factores. En la conformación nativa de 1fjn el potencial detecta hasta cuatro repulsiones (ver Tabla 6.3). Esto dificulta el muestreo en las proximidades de la conformación nativa. Sin embargo, el problema más importante con esta proteína no es de impedimento estérico, sino que procede de las restricciones del potencial de enlace de hidrógeno de Kolinski. En la Figura 6.11 mostramos el mapa de interacciones calculadas con este potencial para la conformación nativa y la de menor energía de 1fjn. En el triángulo superior mostramos interacciones para la conformación nativa. En la zona del mapa de la conformación nativa correspondiente a las dos hebras

(fragmentos 2 y 3) no aparece claramente definido el patrón descrito en el Capítulo 5 para las interacciones entre hebras. Por tanto el potencial no captura las interacciones nativas de la horquilla β. En cambio, si se permutan las posiciones de las hebras, como en la conformación de mínima energía, sí se satisfacen más restricciones (Figura 6.11, triángulo inferior). Esto permite que se calculen los términos energéticos y que se defina un mínimo muy profundo, aunque la contribución hidrófoba pierda peso en la energía global.

La mayoría de las proteínas de tres y cuatro fragmentos que estudiamos (todas menos 1hyw) están estabilizadas por puentes disulfuro entre residuos de cisteína. Para casi todos estos casos, el mínimo energético se encuentra muy próximo a la conformación nativa, aunque el potencial DFIRE-SCM no esté diseñado para describir con precisión este tipo de interacción[32]. La forma en que este potencial incluye el efecto de los puentes disulfuro es un mínimo de -2.35 kcal/mol que tiene entre 2.5 a 3 Å de distancia entre centros de cadenas laterales. Esto es suficiente para identificar como fuertemente atractivos casi todos los puentes disulfuro nativos. En la Figura 6.12 mostramos el mapa de energía para las contribuciones hidrófobas en la conformación nativa y la optimizada de la proteína 1agt. En esta proteína se establecen tres puentes disulfuro, entre los pares de residuos Cys8-Cys28, Cys14-Cys33 y Cys18-Cys35. En la parte correspondiente a la conformación nativa del mapa (triángulo superior) vemos que los tres puentes disulfuro se identifican como interacciones fuertemente atractivas. La conformación de mínima energía alcanzada en la optimización tiene un valor de $RMSD = 1$ Å respecto a la nativa. A pesar de la semejanza entre ambas estructuras, en la minimizada la distancia entre centros de cadenas laterales para algunas de estas cisteínas está levemente desplazada con respecto a la nativa. Por este motivo, en el mapa para la conformación de mínima energía (triángulo inferior en la Figura 6.12) las citadas interacciones entre cisteínas no tienen el valor de energía que corresponde al puente disulfuro. Consideramos muy meri-

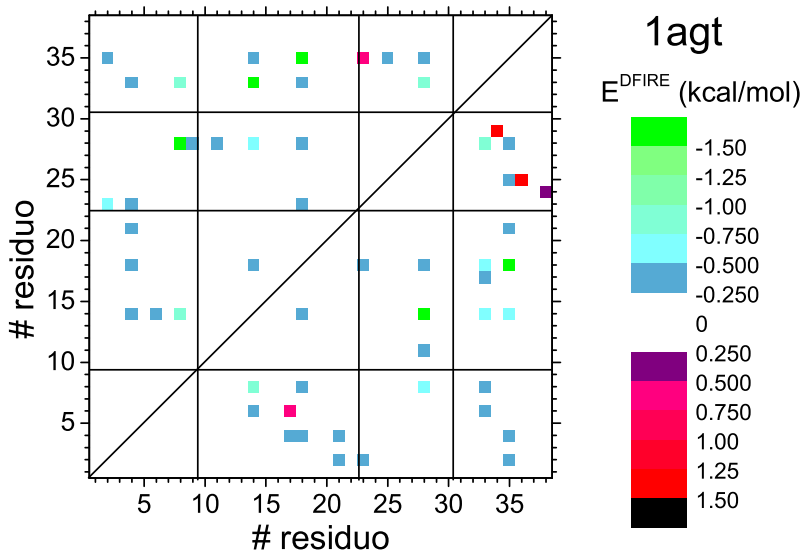


Figura 6.12: Mapa de energías de 1agt calculado con la contribución DFIRE-SCM para las conformaciones nativa (triángulo superior) y optimizada con el potencial DFIRE-Kolinski (triángulo inferior). La energía se expresa en unidades del modelo (kcal/mol).

torio que el potencial DFIRE-SCM sea capaz de definir el núcleo hidrófobo de manera tan robusta como para que el mínimo aparezca en la conformación nativa a pesar de la disminución en la contribución de interacciones específicas los puentes disulfuro. Aun así, este aspecto debe ser tratado con cuidado cuando se emplee este potencial para llevar a cabo simulaciones del plegamiento.

Con la proteína 1c56 sí que observamos la problemática de tratar los puentes disulfuro como una más entre las interacciones entre pares. Para esta proteína el mínimo aparece en una conformación con $RMSD \simeq 2$ Å, muy parecida a la conformación nativa. Pero en este caso la superficie de energía no dirige la búsqueda suavemente hacia este mínimo, como se puede deducir de la correspondiente representación de la Figura 6.9. De hecho, para esta proteína hemos tenido que aumentar hasta quince el número de minimizaciones independientes de cada ejecución del algoritmo evolutivo al observar dificultades en el muestreo. La problemática definición de la superficie de energía se

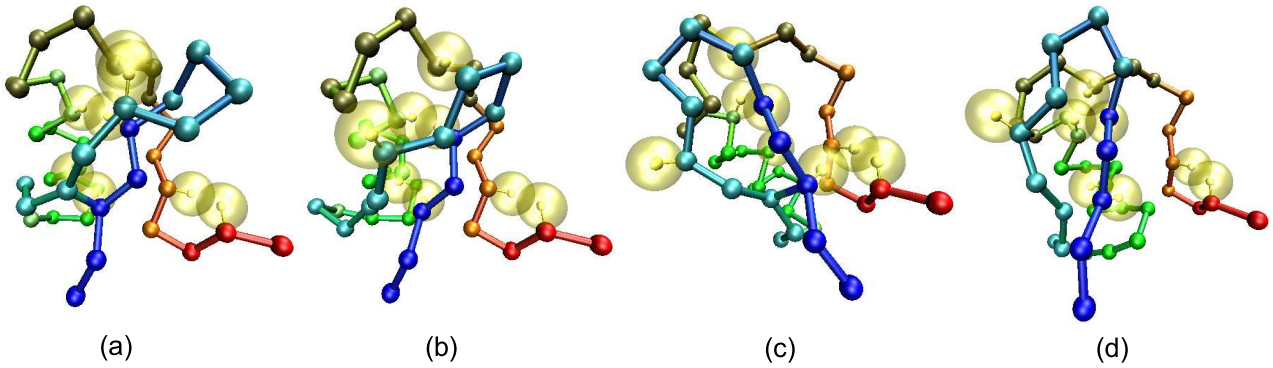


Figura 6.13: Representaciones esquemáticas de distintas conformaciones de la proteína 1c56: (a) nativa, (b) mínimo energético con $RMSD$=2.1 Å, (c) mínimo local con $RMSD$=5.5 Å, y (d) conformación de baja energía con $RMSD$=3.8 Å.

debe, de nuevo, a una serie de factores.

Uno de estos factores es, como hemos comentado, el problema de los puentes disulfuro en el potencial DFIRE-SCM. El patrón de puentes disulfuro de la proteína 1c56 es muy parecido al de 1agt, aunque en 1c56 hay un puente disulfuro más en el extremo de un fragmento del modelo, que no contribuye significativamente a estabilizar su estructura tridimensional[171]. Como en el caso de 1agt, en el mínimo energético alcanzado con nuestro método no todos los pares de cisteínas se encuentran formando sus interacciones nativas. Pero lo que no supone un problema para la definición de la superficie de energía para 1agt, sí lo es para 1c56. Al poder formarse distintas combinaciones entre las cisteínas de la proteína —distintos "puentes disulfuro"— se pueden estabilizar diferentes mínimos locales. En la Figura 6.13 mostramos una serie de representaciones con distintas conformaciones que se obtienen a lo largo de la optimización, con diferentes distribuciones de las cisteínas de la proteína. La primera de las representaciones (a) corresponde a la conformación nativa y la segunda (b), a la optimizada. Como se puede observar las dos conformaciones son muy parecidas, aunque hay diferencias en las interacciones de algunas cisteínas. Las representaciones (c) y (d) corresponden a conformaciones en las que hay tres centros de cadenas laterales de cisteínas formando parte simultáneamente

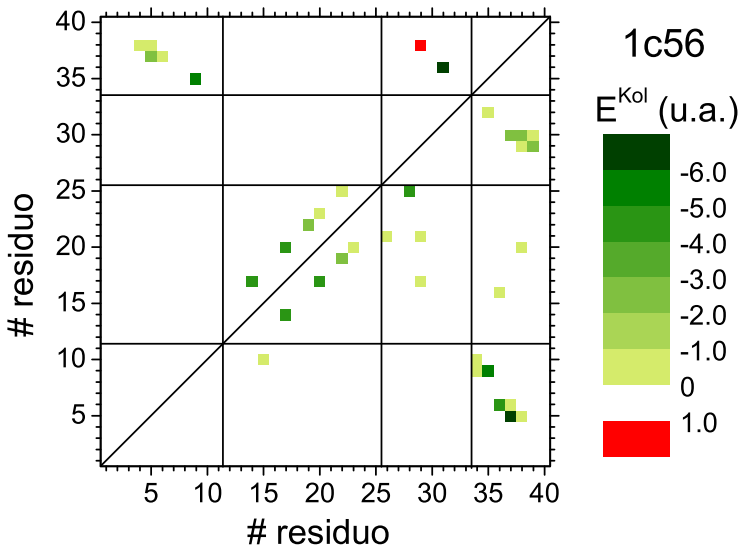


Figura 6.14: Mapa de energías de 1c56 calculado con el potencial de Kolisnki para las conformaciones nativa (triángulo superior) y optimizada con el potencial DFIRE-Kolinski (triángulo inferior). La energía se expresa en unidades del modelo.

de falsos "puentes disulfuro".

La diferencia con 1agt es que la estabilización de estas conformaciones sí supone un conflicto en la minimización de la energía para 1c56. En el caso de 1agt, los potenciales sí identifican el núcleo hidrófobo y los enlaces de hidrógeno de la conformación nativa, de modo que el mínimo nativo es profundo y está bien definido. Así, no importa que se formen ocasionalmente puentes disulfuro no nativos, porque globalmente corresponden a conformaciones menos estables de la proteína. En cambio, en el caso de 1c56 el potencial de Kolinski no es capaz de reconocer los enlaces de hidrógeno nativos. En la Figura 6.14 mostramos el mapa de enlaces de hidrógeno de 1c56 para la conformación nativa y para el mínimo. En ambas partes del mapa, y especialmente en la forma nativa, en la región de interacción entre las hebras (entre los fragmentos 1 y 4, y 3 y 4) apenas aparece el patrón propio de las láminas β. Pensamos que esta mala representación de las interacciones nativas se debe a la poca regularidad de las hebras de la conformación

nativa de 1c56. Según los autores que elucidaron la estructura de esta proteína sólo el 20 % de sus residuos tienen estructura β[171], frente al 47 % de 1agt[172]. Debido a esta falta de regularidad de las regiones tipo β en 1c56, el potencial de Kolinski, ajustado a partir de estructuras regulares de tipo α y β, no propicia una buena definición del mínimo en la conformación nativa. Por el contrario, sí es capaz de definir correctamente el mínimo para 1agt, mucho más regular. Concluimos, por tanto, que la causa de la mala definición del mínimo energético para 1c56 es una conjunción de todos estos factores: la pobre definición de los enlaces de hidrógeno, las disposiciones alternativas de los puentes disulfuro de las cisteínas y las repulsiones de volumen excluido.

La última proteína cuyos resultados analizamos aquí pormenorizadamente es 1hyw, que pertenece a un tipo estructural que no ha sido observado en otras proteínas[173]. Mostramos una representación de su conformación nativa en la Figura 6.15 (a), en la que aparecen las dos hélices α y las dos hebras β que la forman. En este caso, en nuestros experimentos de minimización observamos que hay una cierta degeneración en el mínimo (ver Figura 6.9). Esta degeneración procede del compromiso entre las dos contribuciones a la energía que consideramos. Mostramos representaciones de dos de estos mínimos en la Figura 6.15. Uno de ellos (b), con $RMSD$ frente a la conformación nativa de 3.8 Å, corresponde a un "isómero topológico". El otro mínimo (c), con $RMSD = 2.2$ Å, es muy parecido a la estructura nativa, aunque el solapamiento entre las hélices es algo más eficaz. En el mínimo no nativo (b) la contribución de enlace de hidrógeno recibe más peso que en el mínimo nativo, $E^{Kol}_{(b)} = -26.9$ kcal/mol y $E^{Kol}_{(c)} = -22.1$ kcal/mol. Por el contrario, la contribución hidrófoba es más importante en el mínimo nativo que en el no nativo, $E^{DFIRE}_{(c)} = -34.4$ kcal/mol y $E^{DFIRE}_{(b)} = -30.7$ kcal/mol. De modo que en función de la contribución del potencial que se minimice, el mínimo que se encuentra es diferente.

Para proteínas de mayor tamaño, como 2gb1 y 1j27, el potencial tiene un compor-

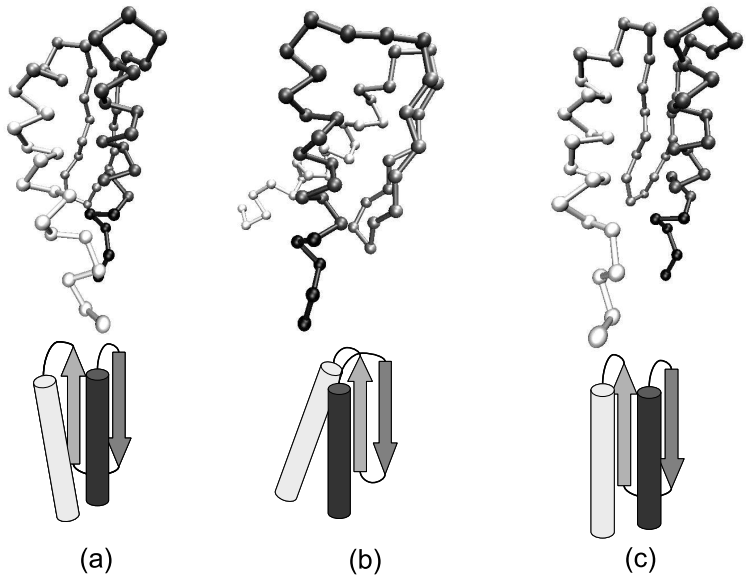


Figura 6.15: Representación de tres conformaciones de la proteína 1hyw: (a) nativa, (b) mínimo con $RMSD$ =3.8 Å, y (c) mínimo con $RMSD$ =2.2 Å, ambos alcanzados en las minimizaciones con el potencial DFIRE-Kolinski.

tamiento muy bueno, a pesar del gran número de interacciones que generan la superficie de energía. Como en el caso de 1hye, en el caso de 1j27 se vuelve a hacer patente el compromiso entre las restricciones del potencial de Kolinski y la definición del núcleo hidrófobo. Para esta proteína la energía de la contribución del potencial DFIRE-SCM disminuye en valor absoluto muy notablemente en la estructura optimizada con respecto a la nativa (ver Tabla 6.3). Pero en este caso este compromiso entre interacciones no induce degeneración en el mínimo, y la combinación de potenciales es capaz de definir un mínimo energético fácilmente accesible en la conformación nativa.

6.5. Resumen del Capítulo y conclusiones

En los Capítulos 4 y 5 hemos estudiado por separado potenciales para la interacción hidrófoba de las cadenas laterales de proteínas y el enlace de hidrógeno del esqueleto. Aquí estudiamos los potenciales que mejores resultados han ofrecido en estas evaluaciones independientes. Para la interacción hidrófoba seleccionamos el potencial DFIRE-SCM, que como hemos visto es capaz de definir el mínimo en la conformación nativa con gran eficacia para proteínas todo α. En el caso de los enlaces de hidrógeno la elección del mejor potencial ha sido más complicada. Nos hemos centrado en los dos potenciales de menor resolución, por parecernos más interesantes para futuros estudios de simulación. Para tener una estimación de cómo funcionarían estos potenciales al combinarlos con el potencial hidrófobo, hemos realizado una serie de cálculos preliminares con nuestro método de minimización. En los resultados de estos experimentos hemos observado que la combinación del potencial de Chen[104] con el DFIRE-SCM (a la que hemos llamado potencial DFIRE-Chen) conserva las buenas propiedades de ambos al utilizarlo con proteínas todo α o todo β. Sin embargo, al utilizarlo con proteínas de tipo $(\alpha + \beta)$ se pierde esta eficacia. El potencial de Chen tiene un efecto de colapso debido a la falta de direccionalidad en su definición de los enlaces de hidrógeno. Esto hace que la combinación de potenciales no tenga la capacidad de definir el mínimo energético eficientemente. En el caso de la combinación del potencial DFIRE-SCM con el de Kolinski (potencial DFIRE-Kolinski) hemos oservado que el potencial sí es capaz de definir el mínimo en la conformación nativa también para proteínas de tipo $(\alpha + \beta)$, por lo que lo hemos estudiado más a fondo.

En ambos casos hemos utilizado una parametrización muy sencilla para dar peso a los dos tipos de interacción. A partir de una serie de aproximaciones, Zhou *et al.* expresan los valores de energía entre pares de su potencial con unidades de energía por

mol[123]. En el caso de los potenciales de enlace de hidrógeno, hemos calculado la máxima contribución que puede alcanzar un enlace de hidrógeno en unidades del modelo, para normalizar los valores de energía calculados. Estos valores se multiplican por la contribución real de un enlace de hidrógeno, entre 3 y 10 kcal/mol[24,25]. Probando distintos valores en este intervalo encontramos que no parece existir una fuerte dependencia con el valor que se le dé al peso de la contribución de enlace de hidrógeno, siempre que se garantice que está suficientemente representado en la energía total.

Para evaluar la combinación de potenciales DFIRE-Kolinski hemos realizado una serie de experimentos de minimización con proteínas de varios tipos. En primer lugar, con proteínas todo α hemos observado que la combinación de potenciales DFIRE-Kolinski conserva las buenas propiedades del potencial DFIRE-SCM para representar la interacción específica entre cadenas laterales inducida por el colapso hidrófobo, como comprobamos en el Capítulo 4. La contribución de Kolinski no participa más que marginalmente en la estabilización global de los empaquetamientos de hélices sobre los que nuestro algoritmo realiza el muestreo. A continuación, hemos realizado una serie de experimentos con proteínas todo β, donde la contribución mayoritaria en nuestro modelo es la de enlace de hidrógeno. En este caso, el potencial hidrófobo contribuye de manera algo más significativa que en el caso de las proteínas todo α. La principal conclusión de esta parte del estudio es que el potencial de Kolinski conserva sus buenas propiedades, descritas en el Capítulo 5. Por su parte, la contribución del potencial DFIRE-SCM reproduce apropiadamente las interacciones entre cadenas laterales por encima y por debajo de las láminas β.

La parte más novedosa de este Capítulo es el estudio de proteínas ($\alpha + \beta$) con la combinación de potenciales DFIRE-Kolinski. Hemos seleccionado un conjunto de proteínas de pequeño tamaño, con hasta 6 fragmentos en nuestro modelo, formadas por una lámina β y una o dos hélices α. Esto nos ha permitido estudiar con detalle cómo los

dos potenciales son capaces de actuar cooperativamente para generar una superficie de energía que, a diferencia de lo que sucedía con la combinación DFIRE-Chen, localice el mínimo energético en la conformación nativa. Así, para la mayoría de las proteínas que hemos estudiado, el potencial define eficazmente una superficie de energía cuyo mínimo, definido en la conformación nativa, se alcanza suavemente en la minimización.

Para las distintas proteínas de tipo $(\alpha + \beta)$ hemos ido estudiando una serie de particularidades que deben ser tenidas en cuenta cuando se utilicen estos potenciales en estudios de simulación del plegamiento. Uno de los aspectos a destacar es que, si bien los dos modelos conservan al ser utilizados conjuntamente sus mayores virtudes, también conservan algunas de sus desventajas. Por ejemplo, hemos observado que las restricciones del potencial de Kolinski pueden contribuir a una mala definición de la superficie de energía. Esto lo hemos observado para las proteínas 1fjn y 1c56. En ambos casos, el potencial no es capaz de detectar los enlaces de hidrógeno que se forman entre hebras en la conformación nativa de la proteína. Por este motivo, en el caso de 1fjn, una proteína en cuya conformación nativa aparecen una horquilla β y una hélice α, el mínimo se define en una conformación en que las hebras de la horquilla tienen permutadas sus posiciones originales. Esto es así porque en esa conformación alternativa el modelo de Kolinski es capaz de identificar más enlaces de hidrógeno que en la horquilla nativa. Como hemos dicho en el Capítulo 5, este tipo de comportamiento puede estar influido por nuestro muestreo sobre un número relativamente pequeño de grados de libertad. Quizás con un método de búsqueda más libre, el potencial sea capaz de generar un mínimo bien definido en una conformación más cercana a la disposición nativa de la horquilla.

También hemos prestado especial atención a la descripción de las interacciones entre cisteínas en el potencial DFIRE-SCM. Como los propios autores señalan, para elaborar su potencial no han hecho distinción entre las cisteínas oxidadas —formando

puentes disulfuro— o reducidas. En nuestro método hemos tratado estas proteínas sin imponer restricciones adicionales. Por ello, hemos detectado la posibilidad de que se formen empaquetamientos alternativos de los residuos de cisteína. Pueden formarse así "falsos puentes disulfuro", por ejemplo entre más de dos cisteínas, que pueden estabilizar artificialmente conformaciones alejadas de la nativa. Lo más interesante con respecto a los puentes disulfuro es que hemos observado una cierta independencia de los resultados respecto a estas interacciones cuando el núcleo hidrófobo y los enlaces de hidrógeno son identificados apropiadamente por el potencial, como en el caso de 1agt.

Un aspecto que resulta muy interesante es cómo interactuan entre sí las dos contribuciones energéticas del potencial. En esta parte del estudio hemos observado que al combinar los potenciales la superficie de energía se vuelve en ocasiones mucho más compleja. Un ejemplo de esta complejidad es la proteína 1c56. Con esta proteína hemos visto que se unen los problemas de volumen excluido entre los centros de interacción del modelo, la definición de los enlaces de hidrógeno y el potencial hidrófobo. Así, las distintas contribuciones parecen enfrentarse unas con otras, por lo que se genera una superficie de energía de gran rugosidad. También en los casos de 1hyw y 1j27 hemos comprobado el efecto de este compromiso entre los distintos tipos de interacción, pero con distinta repercusión en la definición del mínimo energético. Nuestros resultados, además, están influidos por la búsqueda conformacional limitada a empaquetamientos de fragmentos rígidos que hemos realizado. Probablemente, el uso de un modelo de la proteína en el que se considere un mayor número de grados de libertad contribuiría a superar algunas de las dificultades que suponen las restricciones del modelo de Kolinski. Sin embargo, por su eficacia para generar superficies de energía pensamos que la combinación de potenciales DFIRE-Kolinski, quizás con una serie de ajustes menores, podría ser utilizada para estudios de simulación del plegamiento proteínas.

Capítulo 7

Conclusiones generales de esta Tesis

En esta memoria se ha resumido el trabajo que hemos llevado a cabo con el objetivo de evaluar potenciales de interacción para el plegamiento de proteínas. A continuación, enumeramos las principales conclusiones de este estudio.

- Los algoritmos de tipo evolutivo pueden utilizarse como herramienta de búsqueda del mínimo de la superficie de energía de proteínas con una representación reducida de su geometría. Este tipo de método es capaz de alcanzar la conformación de mínima energía rápida y eficazmente, y evita el riesgo de los métodos de simulación molecular tradicionales de que el muestreo quede atrapado en mínimos locales.

- La eficiencia del método evolutivo dependen tanto de la dimensión del problema de búsqueda, es decir, el número de grados de libertad sobre el que se realiza el muestreo, como de la definición de la superficie de energía.

- El aspecto más determinante para que el algoritmo de minimización sea eficaz no es tanto el ajuste de los parámetros genéticos que regulan los operadores de replicación, entrecruzamiento y mutación, como fundamentalmente el mantenimiento de la variabilidad en la población de soluciones. En nuestra metodología esto lo

hemos conseguido utilizando dos estrategias. En primer lugar, llevando a cabo optimizaciones independientes que comparten información en un determinado punto, en el que además parte de la población de soluciones se renueva. En segundo lugar, estableciendo unos criterios de selección de cromosomas de la progenie basados en la diferencia estructural entre individuos de la población.

- La codificación de las soluciones del algoritmo afecta considerablemente a la eficiencia de la metodología. En nuestro caso, la manera en que se traducen las cadenas de variables en conformaciones de la proteína tiene una repercusión directa en los resultados de la búsqueda. También hemos comprobado que la codificación está estrechamente ligada al problema concreto que se trata de resolver. Por este motivo, hemos empleado la codificación más costosa computacionalmente en la mayoría de las aplicaciones del método con potenciales realistas.

- De nuestro estudio sobre potenciales de interacción hidrófoba concluimos que el potencial DFIRE-SCM de Zhou *et al.*[96,123] captura eficazmente las tendencias específicas de los aminoácidos hidrófobos para aparecer empaquetados en las conformaciones nativas de las proteínas. Este potencial representa las interacciones de manera más eficiente que los potenciales de Nanias *et al.*[100] y TE-13, de Tobi y Elber[92,93].

- Del estudio de potenciales de enlace de hidrógeno, concluimos que los modelos sencillos, como el de Chen[104] y el de Kolinski[103], tienen una gran eficiencia para reproducir los distintos tipos de estructura secundaria que aparecen en proteínas. En todo caso, cuando se utiliza este tipo de modelo es necesario tener un cuidado meticuloso en aspectos como el nivel de resolución con que se representa la proteína, el tipo de función para el cálculo de la energía, o las restricciones de volumen excluido. El equilibrio entre estos detalles permitirá o no que el modelo funcione

apropiadamente.

- De nuestro estudio con la combinación de potenciales hidrófobo y de enlace de hidrógeno concluimos que el potencial DFIRE-Kolinski es capaz de representar adecuadamente la superficie de energía de proteínas para estas dos contribuciones. Hemos observado que, al menos con modelos simplificados como el nuestro, el potencial DFIRE-Kolinski es relativamente insensible a la parametrización que determina el peso de cada una de las contribuciones, mientras ambas estén suficientemente representadas en la energía global de las conformaciones compactas.

- El meticuloso análisis de los resultados, basado en la disección individualizada de los diferentes contactos presentes en las estructuras nativas y en las optimizadas, nos ha permitido estudiar en detalle los puntos fuertes y las debilidades de cada uno de los potenciales de interacción. Hemos completado así un estudio sistemático y riguroso cuyos frutos podrán ser aplicados en futuros estudios de simulación sobre el proceso de plegamiento de proteínas.

Bibliografía

1. T. E. Creighton. *Proteins: Structures and Molecular Properties.* W. H. Freeman, Nueva York, 1993.

2. G. J. Kleywegt y T. A. Jones. Phi/psi-chology: Ramachandran revisited. *Structure*, 4:1395–1400, 1996.

3. C. M. Dobson, A. Sali, y M. Karplus. Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Ed.*, 37:868–893, 1998.

4. C. B. Anfinsen. Studies on the principles that govern the folding of protein chains. En T. Frängsmyr y S. Forsén, editores, *Nobel Lectures, Chemistry 1971-1980.* World Scientific Publishing Co., Singapore, 1993.

5. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

6. K. A. Dill y H. S. Chan. From Levinthal to pathways to funnels. *Nature Struct. Biol.*, 4:10–19, 1997.

7. H. S. Chan y K. A. Dill. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins*, 30:2–33, 1998.

8. M. Karplus. The Levinthal paradox: Yesterday and today. *Fold. & Des.*, 1:S69–S75, 1997.

9. C. Levinthal. Are there pathways for protein folding? *J. Chim. Physique*, 65:44–45, 1968.

10. C. Levinthal. How to fold graciously. En P. Debrunner, J. C. M. Tsibris, y E. Münck, editores, *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Monticello, Illinois.* University of Illinois Press, Urbana, 1969.

11. V. S. Pande, A. Y. Grosberg, T. Tanaka, y D. S. Rokhsar. Pathways for protein folding: Is a new view needed? *Curr. Opin. Struc. Biol.*, 8:68–79, 1998.

12. E. Shakhnovich, G. Farztdinov, A. Gutin, y M. Karplus. Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys. Rev. Lett.*, 67:1665–1668, 1991.

13. C. Camacho y D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA*, 90:6369–6372, 1993.

14. A. Sali, E. Shakhnovich, y M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

15. H. S. Chan y K. A. Dill. Transition states and folding of proteins and heteropolymers. *J. Chem. Phys.*, 100:9238–9257, 1994.

16. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, y H. S. Chan. Principles of protein folding –A perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

17. J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, y N. Socci. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA*, 92:3626–3630, 1995.

18. M.-H. Hao y H. A. Scheraga. How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. USA*, 93:4984–4989, 1996.

19. N. Socci, J. N. Onuchic, y P. G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.*, 104:5860–5868, 1996.

20. J. N. Onuchic. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

21. A. R. Fersht, L. S. Itzhaki, N. F. ElMasry, J. M. Matthews, y D. E. Otzen. Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl. Acad. Sci. USA*, 91:10426–10429, 1994.

22. J. N. Onuchic y P. G. Wolynes. Theory of protein folding. *Curr. Opin. Struc. Biol.*, 14:70–75, 2004.

23. C. L. Brooks III, M. Gruebele, J. N. Onuchic, y P. G. Wolynes. Chemical physics of protein folding. *Proc. Natl. Acad. Sci. USA*, 95:11037–11038, 1998.

24. K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

25. C. Gómez-Moreno y J. Sancho. *Estructura de proteínas.* Ariel, Barcelona, 2003.

26. D. F. Stickle, L. G. Presta, K. A. Dill, y G. D. Rose. Hydrogen bonding in globular proteins. *J. Mol. Biol.*, 226:1143–1159, 1992.

27. J. K. Myers y C. N. Pace. Hydrogen bonding stabilizes globular proteins. *Biophys. J.*, 71:2033–2039, 1996.

28. J. Fernández-Recio y A. Romero. Energetics of a hydrogen bond (charged and neutral) and of a cation-π interaction in apoflavodoxin. *J. Mol. Biol.*, 290:319–330, 1999.

29. R. L. Baldwin. In search of the energetic role of peptide hydrogen bonds. *J. Biol. Chem.*, 278:17581–17588, 2003.

30. N. T. Southall, K. A. Dill, y D. J. Haymet. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106:521–533, 2002.

31. G. D. Rose y R. Wolfenden. Hydrogen bonding, hydrophobicity, packing and protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 22:381–415, 1993.

32. H. Zhou y Y. Zhou. Quatifying the effect of burial of amino acid residues on protein stability. *Proteins*, 54:315–322, 2004.

33. Z. S. Hendsch y B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, 3:211–226, 1994.

34. J. M. Sánchez-Ruiz y G. I. Makhatadze. To charge or not to charge? *Trends Biotechnol.*, 19:132–135, 2001.

35. D. Baker y A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.

36. J. Moult. A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struc. Biol.*, 15:285–289, 2005.

37. J. E. Shea y C. L. Brooks III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.

38. A. R. Leach. *Molecular Modelling. Principles and applications.* Prentice-Hall, Harlow, 2001.

39. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, y P. A. Kollman. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

40. A. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. M. S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, y M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

41. C. Oostenbrink, A. Villa, A. E. Mark, y W. F. van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25:1656–1676, 2004.

42. W. L. Jorgensen, D. S. Maxwell, y J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

43. M. P. Allen y D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Science, Oxford 1987.

44. D. Frenkel y B. Smit. *Understanding Molecular Simulation: From algorithms to applications*. 2ª Ed. Academic Press, San Diego, 2002.

45. D. A. C. Beck y V. Daggett. Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods*, 34:112–120, 2004.

46. M. Karplus y J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Struct. Biol.*, 9:646–652, 2002.

47. J. Hermans. Hydrogen bonds in molecular mechanics force fields. *Adv. Protein. Chem.*, 72:105–119, 2005.

48. J. Kubelka, J. Hofrichter, y W. A. Eaton. The protein folding 'speed limit'. *Curr. Opin. Struc. Biol.*, 14:76–88, 2004.

49. L. Mirny y E. Shakhnovich. Protein folding theory: From lattice to all atom models. *Annu. Rev. Biophys. Biomol. Struct.*, 30:361–396, 2001.

50. A. Kolinski y J. Skolnick. Reduced models of proteins. *Polymer*, 45:511–524, 2004.

51. V. Tozzini. Coarse-grained models for proteins. *Curr. Opin. Struc. Biol.*, 15:144–150, 2005.

52. A. Kolinski y J. Skolnick. Assembly of protein structure from sparse experimental data: An efficient Monte Carlo model. *Proteins*, 32:475–494, 1998.

53. J. Pillardy, C. Czaplewski, A. Liwo, W. J. Wedemeyer, J. Lee, D. R. Ripoll, P. Arlukowicz, S. Oldziej, Y. A. Arnautova, y H. A. Scheraga. Development of physics-based energy functions that predict medium-resolution structures for proteins of the α, β and α/β structural classes. *J. Phys. Chem. B*, 105:7299–7311, 2001.

54. R. L. Jernigan e I. Bahar. Structure-derived potentials and protein simulations. *Curr. Opin. Struc. Biol.*, 6:195–209, 1996.

55. J. Skolnick. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struc. Biol.*, 16:166–171, 2006.

56. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, y P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

57. N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. K. Green, J. L. Flippen-Anderson, J. Westbrook, H. M. Berman, y P. E. Bourne. The RCSB Protein Data Bank: A redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, 33:D233–D237, 2005.

58. H. A. Carlson y J. A. McCammon. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.*, 57:213–218, 2000.

59. L. W. Lee y J.-S. Wang. Flat histogram simulation of lattice polymer systems. *Phys. Rev. E*, 64:056112(1–7), 2001.

60. F. Liang y W. H. Wong. Evolutionary Monte Carlo for protein folding simulations. *J. Chem. Phys.*, 115:3374–3380, 2001.

61. M. S. Apaydin, D. L. Brutlag, C. Guestrin, D. Hsu, y J.-C. Latombe. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. *J. Comput. Biol.*, 10:257–281, 2003.

62. M. Zhang, R. A. White, L. Wang, R. Goldman, L. Kavraki, y B. Hassett. Improving conformational searches by geometric screening. *Bioinformatics*, 21:624–630, 2005.

63. U. H. Hansmann y Y. Okamoto. New Monte Carlo algorithms for protein folding. *Curr. Opin. Struc. Biol.*, 9:177–183, 1999.

64. D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning.* Addison-Wesley, Reading, 1989.

65. Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* 3Âª Ed. Springer, Berlín, 1999.

66. L. Chambers. *The Practical Handbook of Genetic Algorithms.* Chapman & Hall, Boca Raton, 2001.

67. D. E. Clark y D. R. Westhead. Evolutionary algorithms in computer aided molecular design. *J. Comput. Aid. Mol. Des.*, 10:337–358, 1996.

68. R. Unger. The genetic algorithm approach to protein structure prediction. *Struct. Bond.*, 110:153–175, 2004.

69. J. R. Gunn. Sampling protein conformations using segment libraries and a genetic algorithm. *J. Chem. Phys.*, 106:4270–4281, 1997.

70. G. A. Cox, T. V. Mortimer-Jones, R. P. Taylor, y R. L. Johnston. Development and optimisation of a novel genetic algorithm for studying model protein folding. *Theor. Chem. Acc.*, 112:163–178, 2004.

71. T. Jiang, Q. Cui, C. Shi, y S. Ma. Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. *J. Chem. Phys.*, 119:4592–4596, 2003.

72. C. A. Del Carpio. A parallel genetic algorithm for polypeptide three dimensional structure prediction. A transputer implementation. *J. Chem. Inf. Comput. Sci.*, 36:258–269, 1996.

73. R. Unger y J. Moult. Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231:75–81, 1993.

74. T. Dandekar y P. Argos. Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.*, 236:844–861, 1994.

75. S. M. Le Grand y K. M. Merz Jr. The application of the genetic algorithm to the minimization of potential energy functions. *J. Global Optim.*, 3:49–66, 1993.

76. J. T. Pedersen y J. Moult. Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.*, 269:240–259, 1997.

77. T. Dandekar. Improving protein structure prediction by new strategies: Experimental insights and the genetic algorithm. *J. Mol. Model.*, 3:312–314, 1997.

78. A. A. Rabow y H. A. Scheraga. Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Sci.*, 5:1800–1815, 1996.

79. T. Dandekar y P. Argos. Applying experimental data to protein fold prediction with the genetic algorithm. *Protein Eng.*, 10:877–893, 1997.

80. S. Miyazawa y R. L. Jernigan. Estimation of interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

81. M. J. Sippl. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

82. M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, y M. Sippl. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.*, 216:167–180, 1990.

83. D. Jones, W. Taylor, y J. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

84. V. N. Maiorov y G. M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227:876–888, 1992.

85. M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aid. Mol. Des.*, 7:473–501, 1993.

86. J. Bowie y D. Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA*, 91:4436–4440, 1994.

87. M. J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struc. Biol.*, 5:229–235, 1995.

88. S. Miyazawa y R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

89. M. Vendruscolo, R. Najmanovich, y E. Domany. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins*, 38:134–148, 2000.

90. A. Panchenko, A. Marchler-Bauer, y S. Bryant. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.*, 296:1319–1331, 2000.

91. M. Vijayakumar y H. Zhou. Prediction of residue-residue pair frequencies in proteins. *J. Phys. Chem. B*, 104:9755–9764, 2000.

92. D. Tobi, G. Shafran, N. Linial, y R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40:71–85, 2000.

93. D. Tobi y R. Elber. Distance dependent, pair potential for protein folding: Results from linear optimization. *Proteins*, 41:40–46, 2000.

94. F. Melo y R. Sánchez. Statistical potentials for fold assessment. *Protein Sci.*, 11:430–448, 2002.

95. M. Chhajer y G. M. Crippen. A protein folding potential that places the native states of a large number of proteins near a local minimum. *BMC Struct. Biol.*, 2:4, 2002.

96. C. Zhang, S. Liu, H. Zhou, y Y. Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.*, 13:400–411, 2004.

97. B. H. Park, E. S. Huang, y M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, 266:831–846, 1997.

98. H. Taketomi, Y. Ueda, y N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.*, 7:445–459, 1975.

99. D. de Sancho, L. Prieto, A. M. Rubio, y A. Rey. Evolutionary method for the assembly of rigid protein fragments. *J. Comput. Chem.*, 26:131–141, 2005.

100. M. Nanias, M. Chinchio, J. Pillardy, D. R. Ripoll, y H. A. Scheraga. Packing helices in proteins by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA*, 100:1706–1710, 2003.

101. D. de Sancho y A. Rey. Assessment of protein folding potentials with an evolutionary method. *J. Chem. Phys.*, 125:014904 (1–9), 2006.

102. A. Irbäck, F. Sjunnesson, y S. Wallin. Three-helix-bundle protein in a Ramachandran model. *Proc. Natl. Acad. Sci. USA*, 97:13614–13618, 2000.

103. A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, 51:349–371, 2004.

104. H. Imamura y J. Z. Y. Chen. Dependence of folding dynamics and structural stability on the location of a hydrophobic pair in β-hairpins. *Proteins*, 63:555–570, 2006.

105. D. de Sancho y A. Rey. Evaluation of coarse grained models for hydrogen bonds in proteins. *J. Comput. Chem.*, 28:1187–1199 2007.

106. D. T. Jones. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, Suppl. 1:185–191, 1997.

107. B. Fain y M. Levitt. A novel method for sampling alpha-helical protein backbones. *J. Mol. Biol.*, 305:191–201, 2001.

108. G. Chikenji, Y. Fujitsuka, y S. Takada. A reversible fragment assembly method for *de novo* protein structure prediction. *J. Chem. Phys.*, 119:6895–6903, 2003.

109. J. Skolnick, Y. Zhang, A. K. Arakaki, A. Kolinski, M. Boniecki, A. Szilági, y D. Kihara. TOUCHSTONE: A unified approach to protein structure prediction. *Proteins*, 53:469–479, 2003.

110. N. Haspel, C.-J. Tsai, H. Wolfson, y R. Nussinov. Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci.*, 12:1177–1187, 2003.

111. T. X. Hoang, F. Seno, J. R. Banavar, M. Cieplak, y A. Maritan. Assembly of protein tertiary structures from secondary structures using optimized potentials. *Proteins*, 52:155–165, 2003.

112. J. Lee, S.-Y. Kim, K. Joo, I. Kim, y J. Lee. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins*, 56:704–714, 2004.

113. Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, y P. G. Wolynes. Optimizing physical energy functions for protein folding. *Proteins*, 54:88–103, 2004.

114. P. Tuffery y P. Derreumaux. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Gō potential and a greedy algorithm. *Proteins*, 61:732–740, 2005.

115. C. Cohen y D. A. D. Parry. α-helical coiled coils —A widespread motif in proteins. *Trends Biochem. Sci.*, 11:245–248, 1986.

116. R. Brüschweiler. Efficient RMSD measures for the comparison of two molecular ensembles. *Proteins*, 50:26–34, 2003.

117. D. Whitley. An overview of evolutionary algorithms: Practical issues and common pitfalls. *Inform. Software Technology*, 43:817–831, 2001.

118. E. Paci, M. Vendruscolo, y M. Karplus. Validity of Gō models: Comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.*, 83:3032–3038, 2002.

119. G. Settanni, T. X. Hoang, C. Micheletti, y A. Maritan. Folding pathways of prion doppel. *Biophys. J.*, 83:3533–3541, 2002.

120. J. Tsai, R. Taylor, C. Chothia, y M. Gerstein. The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.*, 290:253–266, 1999.

121. L. Prieto, D. de Sancho, y A. Rey. Thermodynamics of Gō-type models for protein folding. *J. Chem. Phys.*, 123:154903 (1–8), 2005.

122. T. L. Hill. *Statistical Mechanics: Principles and selected applications.* Courier Dover, Nueva York, 1956.

123. H. Zhou y Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, 11:2714–2726, 2002.

124. D. Frishman y P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23:566–579, 1995.

125. C. L. Brooks III. Simulations of protein folding and unfolding. *Curr. Opin. Struc. Biol.*, 1998:222–226, 1998.

126. P. J. Flory. *Statistical mechanics of chain molecules.* Interscience, Nueva York, 1969.

127. P. Pokarowski, A. Kloczkowski, R. L. Jernigan, N. S. Kothari, M. Pokarowska, y A. Kolinski. Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins*, 59:49–57, 2005.

128. S. Miyazawa y R. L. Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, 34:49–68, 1999.

129. J. Skolnick, A. Kolinski, y A. R. Ortiz. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, 265:217–241, 1997.

130. D. Hinds y M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668–682, 1994.

131. A. Godzik. Knowledge-based potentials for protein folding: What can we learn from known protein structures? *Structure*, 4:363–366, 1996.

132. M. R. Betancourt y D. Thirumalai. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.*, 8:361–369, 1999.

133. I. Bahar y R. L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266:195–214, 1997.

134. H. Lu y J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44:223–232, 2001.

135. R. Samudrala y J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916, 1998.

136. H. Li y Y. Zhou. Fold helical proteins by energy minimization in the dihedral space and a DFIRE-based statistical energy function. *J. Bioinfo. Comput. Biol.*, 3:1151–1170, 2005.

137. S. Liu, C. Zhang, H. Zhou, y Y. Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, 56:93–101, 2004.

138. A. V. Morozov y T. Kortemme. Potential functions for hydrogen bonds in protein structure prediction and design. *Adv. Protein. Chem.*, 72:1–38, 2005.

139. A. E. Mirsky y L. Pauling. On the structure of native, denatured, and coagulated proteins. *Proc. Natl. Acad. Sci. USA*, 22:439–447, 1936.

140. L. Pauling y R. B. Corey. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 37:235–240, 1951.

141. L. Pauling y R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 37:251–256, 1951.

142. L. Pauling y R. B. Corey. The structure of fibrous proteins of the collagen-gelatin group. *Proc. Natl. Acad. Sci. USA*, 37:272–281, 1951.

143. L. Pauling y R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc. Natl. Acad. Sci. USA*, 37:729–740, 1951.

144. B. Zagrovic, C. D. Snow, M. R. Shirts, y V. S. Pande. Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.*, 323:927–937, 2002.

145. A. Irbäck, F. Sjunnesson, y S. Wallin. Hydrogen bond, hydrophobicity forces and the character of the collapse transition. *J. Biol. Phys.*, 27:169–179, 2001.

146. G. Favrin, A. Irbäck, y S. Wallin. Folding of a small protein using hydrogen bonds and hydrophobicity forces. *Proteins*, 47:99–105, 2002.

147. G. Favrin, A. Irbäck, B. Samuelson, y S. Wallin. Two-state folding over a weak free-energy barrier. *Biophys. J.*, 85:1457–1465, 2003.

148. A. Irbäck, B. Samuelsson, F. Sjunnesson, y S. Wallin. Thermodynamics of α and β-structure formation in proteins. *Biophys. J.*, 85:1466–1473, 2003.

149. A. Irbäck. Protein folding in the absence of a clear free-energy barrier. *Acta Phys. Pol.*, 34:4867–4878, 2003.

150. A. Irbäck, B. Samuelsson, F. Sjunnesson, y S. Wallin. A minimalistic all-atom approach to protein folding. *J. Phys.: Condens. Matter*, 15:1797–1807, 2003.

151. G. Favrin, A. Irbäck, y S. Wallin. Sequence-based study of two related proteins with different folding behaviors. *Proteins*, 54:8–12, 2004.

152. A. Irbäck y F. Sjunnesson. Folding thermodynamics of three β-sheet peptides: A model study. *Proteins*, 56:110–116, 2004.

153. A. Irbäck y S. Mohanty. Folding thermodynamics of peptides. *Biophys. J.*, 88:1560–1569, 2005.

154. A. Irbäck, S. Mitternacht, y S. Mohanty. Dissecting the mechanical unfolding of ubiquitin. *Proc. Natl. Acad. Sci. USA*, 102:13427–13432, 2005.

155. A. Irbäck. Peptide folding and aggregation studied using a simplified atomic model. *J. Phys.: Condens. Matter*, 17:S1553–S1564, 2005.

156. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, y P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

157. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, y M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

158. J. Z. Y. Chen y H. Imamura. Universal model for α-helix and β-sheet structures in proteins. *Physica A*, 321:181–188, 2003.

159. J. Z. Y. Chen, A. S. Lemak, J. R. Lepock, y J. P. Kemp. Minimal model for studying prion-like folding pathways. *Proteins*, 51:283–288, 2003.

160. H. Imamura y J. Z. Y. Chen. Conformational conversion of proteins due to mutation. *Europhys. Lett.*, 67:491–497, 2004.

161. H. Imamura y J. Z. Y. Chen. Minimum model for the α-helix-β-hairpin transition in proteins. *Proteins*, 67:459–468, 2007.

162. Y. Zhang, A. Kolinski, y J. Skolnick. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.*, 85:1145–1164, 2003.

163. D. Kihara, H. Lu, A. Kolinski, y J. Skolnick. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA*, 98:10125–10130, 2001.

164. M. Boniecki, P. Rotkiewicz, S. Jeffrey, y A. Kolinski. Protein fragment reconstruction using various modeling techniques. *J. Comput. Aid. Mol. Des.*, 17:725–738, 2003.

165. D. Ekonomiuk, M. Kielbasinski, y A. Kolinski. Protein modeling with reduced representation: Statistical potentials and protein folding mechanism. *Acta Biochim. Pol.*, 52:741–758, 2005.

166. D. Plewczynska y A. Kolinski. Protein folding with a reduced representation and inaccurate short-range restraints. *Macromol. Theor. Simul.*, 14:444–451, 2005.

167. S. Kmiecik, M. Kurcinski, A. Rutkowska, D. Gront, y A. Kolinski. Denatured proteins and early folding intermediates simulated in a reduced conformational space. *Acta Biochim. Pol.*, 53:131–143, 2006.

168. W. Kabsch y C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

169. C. M. Wilmot y J. M. Thornton. Analysis and prediction of the different types of β turns in proteins. *J. Mol. Biol.*, 203:221–232, 1988.

170. C. M. Wilmot y J. M. Thornton. β-turns and their distortions: A proposed new nomenclature. *Protein Eng.*, 3:479–493, 1990.

171. S. K. Holaday Jr., B. M. Martin, P. L. Fletcher Jr., y N. R. Krishna. NMR solution structure of butantoxin. *Arch. Biochem. Biophys.*, 379:18–27, 2000.

172. A. M. Krezel, C. Kasibhatla, P. Hidalgo, R. MacKinnon, y G. Wagner. Solution structure of the potassium channel inhibitor agitoxin 2: Caliper for probing channel geometry. *Protein Sci.*, 4:1478–1489, 1995.

173. K. L. Maxwell, A. A. Yee, V. Booth, C. H. Arrowsmith, M. Gold, y A. R. Davidson. The solution structure of bacteriophage λ protein W, a small morphogenetic protein possessing a novel fold. *J. Mol. Biol.*, 308:9–14, 2001.